



(51) Classification internationale des brevets :
H04L 29/06 (2006.01) *H04L 12/26* (2006.01)
H04L 29/08 (2006.01)

(21) Numéro de la demande internationale :
PCT/FR2011/051153

(22) Date de dépôt international :
20 mai 2011 (20.05.2011)

(25) Langue de dépôt : français

(26) Langue de publication : français

(30) Données relatives à la priorité :
10 02132 20 mai 2010 (20.05.2010) FR

(71) Déposant (pour tous les États désignés sauf US) :
THALES [FR/FR]; 45 rue de Villiers, F-92200 Neuilly
Sur Seine (FR).

(72) Inventeurs; et

(75) Inventeurs/Déposants (pour US seulement) :
CRAPELLA, Gregory [FR/FR]; 160 boulevard de
Valmy BP 82, F-92704 Colombes Cedex (FR).
BAZELLE, Thibaud [FR/FR]; 160, boulevard de Valmy
BP 82, F-92704 Colombes Cedex (FR). **CHOLLON,
Laurent** [FR/FR]; 160, boulevard de Valmy BP 82,
F-92704 Colombes Cedex (FR).

(74) Mandataires : **BLOT, Philippe** et al.; Cabinet Lavoix, 2,
place d'Estienne d'Orves, F-75009 Paris (FR).

(81) États désignés (sauf indication contraire, pour tout titre
de protection nationale disponible) : AE, AG, AL, AM,
AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ,
CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO,
DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT,
HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP,
KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD,
ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI,
NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, SC, SD,
SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR,
TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) États désignés (sauf indication contraire, pour tout titre
de protection régionale disponible) : ARIPO (BW, GH,
GM, KE, LR, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG,
ZM, ZW), eurasien (AM, AZ, BY, KG, KZ, MD, RU, TJ,
TM), européen (AL, AT, BE, BG, CH, CY, CZ, DE, DK,
EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU,
LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK,
SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ,
GW, ML, MR, NE, SN, TD, TG).

Déclarations en vertu de la règle 4.17 :

— relative à la qualité d'inventeur (règle 4.17.iv))

[Suite sur la page suivante]

(54) Title : METHOD AND DEVICE FOR ANALYSING DATA INTERCEPTED ON AN IP NETWORK IN ORDER TO MONITOR THE ACTIVITY OF WEB SITE USERS

(54) Titre : PROCÉDÉ ET DISPOSITIF D'ANALYSE DE DONNÉES INTERCEPTÉES SUR UN RÉSEAU IP POUR LA SURVEILLANCE DE L'ACTIVITÉ DES UTILISATEURS D'UN SITE WEB

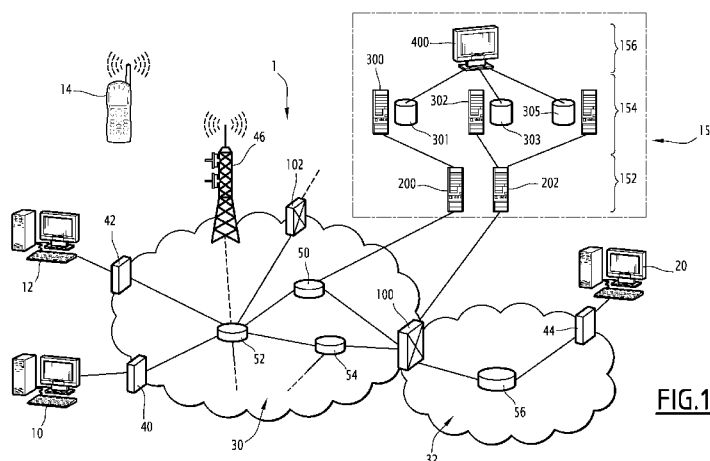


FIG.1

(57) Abstract : The method comprises the following steps consisting in: acquiring (612) the complete data frame from a HTTP request; selecting (614) the data frame acquired if the binary structure thereof meets a plurality of conditions comprising at least one condition corresponding to the IP layer (1040, 1050) of the frame, at least one condition corresponding to the transport layer (1080) of the frame and at least one condition corresponding to the application layer (1100, 1110) of the frame; extracting (616) data of interest from the application layer of the selected frame; and recording (618, 620, 712, 714, 716, 718) the extracted data in a database (301, 303, 305).

(57) Abrégé :

[Suite sur la page suivante]



Publiée :

— avec rapport de recherche internationale (Art. 21(3))

Ce procédé comporte les étapes consistant à: acquérir (612) la trame de données complète d'une requête HTTP; sélectionner (614) la trame de données acquise si sa structure binaire vérifie une pluralité de conditions comportant au moins une condition correspondant à la couche IP (1040, 1050) de la trame, au moins une condition correspondant à la couche de transport (1080) de la trame et au moins une condition correspondant à la couche d'application (1100, 1110) de la trame; extraire (616) des données d'intérêt de la couche d'application de la trame sélectionnée; enregistrer (618, 620, 712, 714, 716, 718) les données extraites, dans une base de données (301, 303, 305).

Procédé et dispositif d'analyse de données interceptées sur un réseau IP pour la surveillance de l'activité des utilisateurs d'un site Web

L'invention a pour domaine celui des procédés et des dispositifs de traitement en temps réel d'un flux de données intercepté sur un réseau de communication IP.

Pour la surveillance d'un site Web particulier, l'administration légalement autorisée (notée ALA dans le document) d'un état se font remettre, par l'hébergeur du site Web ou son administrateur, un ou plusieurs fichiers « log », regroupant l'historique de la connexion sur le serveur d'accès au site Web.

Cette manière de faire implique de mettre l'hébergeur ou l'administrateur au courant du fait que le site Web qu'il héberge est sous écoute.

De plus, si l'hébergeur ou l'administrateur ne relève pas de la loi nationale, le site Web étant hébergé à l'étranger bien que les utilisateurs de ce site Web soient des nationaux de cet état, il est difficile pour l'ALA de contraindre l'hébergeur ou l'administrateur étranger de fournir les fichiers « log ».

L'invention a donc pour but de pallier aux problèmes précités en proposant un procédé et un dispositif d'analyse permettant de traiter, en temps réel, un flux de données interceptées sur un réseau de communication IP pour la surveillance détaillée de l'activité des utilisateurs d'un site Web d'intérêt.

Pour cela l'invention a pour objet un procédé d'analyse de requête HTTP interceptées sur un réseau IP pour la surveillance de l'activité des utilisateurs d'un site Web prédéterminé, comportant les étapes consistant à :

- acquérir la trame de données complète d'une requête HTTP;
- sélectionner la trame de données acquise si sa structure binaire vérifie une pluralité de conditions comportant au moins une condition correspondant à la couche IP de la trame, au moins une condition correspondant à la couche de transport de la trame et au moins une condition correspondant à la couche d'application de la trame;
- extraire des données d'intérêt de la couche d'application de la trame sélectionnée ;
- enregistrer les données extraites, dans une base de données.

Suivant des modes particuliers de réalisation, le procédé comporte une ou plusieurs des caractéristiques suivantes, prises isolément ou suivant toutes les combinaisons techniquement possibles :

- l'étape de sélection permet la sélection d'une trame dont la couche de transport est une couche TCP et dont la couche d'application est une couche HTTP.

- à l'étape de sélection, ladite au moins une condition sur la couche IP, respectivement ladite au moins une condition sur la couche TCP, consiste à comparer la longueur d'un paquet de bits inclus dans la trame acquise, paquet considéré comme étant un paquet IP, respectivement un paquet TCP, avec une longueur prédéfinie de l'en-tête d'un paquet IP, respectivement d'un paquet TCP.

- à l'étape de sélection, ladite au moins une condition sur la couche IP, respectivement ladite au moins une condition sur la couche HTTP, consiste à appliquer, sur l'en-tête d'un paquet de bits inclus dans la trame acquise, paquet considéré comme étant un paquet IP, respectivement un paquet HTTP, un masque pour extraire un groupe de bits et à comparer ce groupe de bits avec valeur binaire attendue pour un paramètre présent dans l'en-tête d'un paquet IP, respectivement l'en-tête d'un paquet HTTP.

- entre l'étape consistant à extraire les données de la couche d'application de ladite trame et à enregistrer ces données dans une base de données, le procédé comporte une étape supplémentaire consistant à mettre en forme les données extraites selon un modèle prédéterminé, de préférence en y associant des meta-données.

L'invention a également pour objet un dispositif pour la mise en œuvre du procédé selon l'une quelconque des revendications 1 à 5, caractérisé en ce qu'il comporte :

- un moyen d'acquisition d'une trame de données complète d'une requête HTTP interceptée sur un réseau de communication IP auquel ledit dispositif est connecté ;
- un moyen de sélection propre à vérifier une pluralité de conditions sur la structure binaire d'une trame de données acquise obtenue en sortie du moyen d'acquisition, et comportant au moins une routine de vérification d'une condition correspondant à la couche IP de la trame, au moins une routine de vérification d'une condition correspondant à la couche de transport de la trame et au moins une routine de vérification d'une condition correspondant à la couche d'application de la trame ;
- un moyen d'extraction apte à extraire des données de la couche d'application d'une trame de données sélectionnée obtenue en sortie du moyen de sélection ;
- un moyen d'enregistrement apte à stocker les données extraites obtenues en sortie du module d'extraction dans une base de données.

Suivant des modes particuliers de réalisation, le dispositif comporte une ou plusieurs des caractéristiques suivantes, prises isolément ou suivant toutes les combinaisons techniquement possibles.

- le moyen de sélection est adapté pour la sélection d'une trame de données acquises dont la couche de transport est une couche TCP et dont la couche d'application est une couche HTTP.

- le dispositif comporte un étage de traitement comportant une pluralité d'ordinateurs serveur de traitement, chaque ordinateur serveur de traitement étant connecté audit réseau de communication IP et comportant une instanciation desdits moyens d'acquisition, de sélection et d'extraction.

5 - le dispositif comporte, en outre, un étage de stockage comportant une pluralité d'ordinateurs serveur de stockage, chaque ordinateur serveur de stockage étant connecté à ladite pluralité d'ordinateurs serveur de traitement, étant associé à au moins une base de données, et comportant une instanciation desdits moyens de stockage apte à stocker les données extraites communiquées par un ordinateur serveur de traitement dans la
10 base de données associée à l'ordinateur serveur de stockage considéré.

- le dispositif comporte, en outre, un étage de restitution comportant au moins un ordinateur de restitution comportant des moyens d'interrogation des différentes bases de données de l'étage de stockage.

La modularité du dispositif, c'est à dire la séparation en modules des étapes de
15 traitement, de stockage et de restitution, et l'extensibilité du dispositif, c'est-à-dire la possibilité d'avoir plusieurs instances de chaque module, autorise l'analyse, en temps réel, d'un flux de données IP présentant un débit très élevé et/ou un volume important.

Grâce à la mise en œuvre d'une étape de sélection comportant une analyse « en
20 profondeur » des données IP incidentes, c'est-à-dire une analyse au niveau binaire des trames, le procédé permet de traiter, en temps réel, un flux de données présentant un débit très élevé, de l'ordre de quelques Gbits. L'étape d'extraction des données présentant un intérêt pour la surveillance du site Web n'est réalisée qu'en aval de l'étape de sélection, sur un nombre réduit de trames sélectionnées.

L'invention et ses avantages seront mieux compris à la lecture de la description
25 qui va suivre, donnée uniquement à titre d'exemple et faite en se référant aux dessins annexés sur lesquels :

- la figure 1 est une représentation schématique de l'architecture matérielle pour la mise en œuvre du procédé de traitement ;

- la figure 2 est une représentation schématique des différents logiciels permettant
30 la mise en œuvre du procédé de traitement ;

- la figure 3 est un organigramme schématique représentant les différentes étapes du procédé d'analyse ;

- la figure 4 est un organigramme détaillé représentant l'étape de filtrage du procédé de traitement ; et,

35 - la figure 5 représente les différentes couches d'une trame.

De manière générique, un ordinateur comporte des moyens de mémorisation, tels qu'une mémoire vive RAM, une mémoire morte ROM et un espace de stockage tels qu'un ou plusieurs disques durs, et des moyens de calcul, tels qu'un processeur, aptes à exécuter les instructions de programmes informatiques qui sont stockées dans les moyens de mémorisation de l'ordinateur.

Un ordinateur comporte en outre des interfaces entrée/sortie adaptées pour la connexion de l'ordinateur à au moins un réseau lui permettant d'entrer en communication avec au moins un autre ordinateur connecté à ce réseau.

En se référant à la figure 1, l'architecture 1 comporte un premier ordinateur client 10, un second ordinateur client 12, et un troisième ordinateur client 14. Les ordinateurs client 10 et 12 sont du type ordinateur personnel (PC) et l'ordinateur client 14 est du type téléphone mobile capable de se connecter à un réseau de téléphonie cellulaire du type 3G.

L'architecture 1 comporte également un ordinateur serveur 20 constituant un serveur HTTP ou Web. Il héberge un site Web à surveiller.

L'architecture 1 comporte deux réseaux de communication IP. Le premier réseau 30 est un réseau géré par un fournisseur d'accès à Internet pouvant coopérer avec l'ALA. Le second réseau 32 est géré par un autre opérateur. Le serveur 20 est connecté au second réseau. En variante, il appartient au premier réseau.

Les réseaux 30 et 32 autorisent la communication IP entre un ordinateur client 10, 12, 14 et le serveur HTTP 20. Les réseaux comportent une pluralité d'équipements d'accès 40, 42, 44 et 46 ainsi qu'une pluralité d'équipements routeur 50, 52 et 54, et d'équipement d'interconnexion entre réseaux 100 et 102.

Un équipement routeur est apte à réémettre un paquet IP incident vers un nœud du réseau que l'équipement routeur choisit en fonction de l'adresse du destinataire final du paquet, adresse que l'équipement routeur est apte à lire dans le paquet incident.

Un équipement d'interconnexion constitue un point d'accès au réseau 30 pour les autres réseaux. Les équipements d'interconnexion 100, 102 sont gérés par le fournisseur d'accès, en accord avec le ou les autres opérateurs des autres réseaux.

Un ordinateur client appartenant à un utilisateur ayant un abonnement auprès du fournisseur d'accès peut être connecté au premier réseau 30 de différentes manières. Ainsi, l'ordinateur client 10 est connecté à l'équipement d'accès 40 par un lien du type ADSL. L'ordinateur 12 est connecté à l'équipement d'accès 42 par un lien du type RTC. Le téléphone mobile 14 est connecté par un lien hertzien à l'équipement d'accès 46. Une adresse IP est attribuée à l'ordinateur client lorsqu'il se connecte à l'équipement d'accès.

Le dispositif pour la mise en œuvre du procédé de traitement est représenté sur la figure 1 et est indiqué par la référence générale 150.

Le dispositif 150 comporte un premier étage de traitement 152. Sur la figure 1, l'étage de traitement comporte deux ordinateurs serveur de traitement 200 et 202.

5 Un serveur de traitement comporte un espace de mémoire adressable.

Un serveur de traitement est connecté, en amont, au premier réseau IP. Ainsi le premier ordinateur de traitement 200 est connecté au routeur 50 et le second ordinateur de traitement 202 est connecté à l'équipement d'interconnexion 100.

10 Un serveur de traitement est connecté, en aval, à un ou plusieurs serveurs de stockage qui vont maintenant être décrits.

Le dispositif 150 comporte un second étage de stockage 154. Sur la figure 1, l'étage de stockage comporte trois ordinateurs serveur de stockage 300, 302 et 304. Chaque serveur de stockage est associé à une base de données, respectivement 301, 303, 305.

15 Enfin, le dispositif 150 comporte un étage de restitution 156. Sur la figure 1, l'étage de restitution comporte un ordinateur client de restitution 400. L'ordinateur client de restitution est connecté à chacune des bases de données 301, 303, 305.

20 Un logiciel d'interception passive est stocké et exécuté sur un ou plusieurs des équipements du premier réseau géré par le fournisseur d'accès. Par exemple, l'équipement d'interconnexion 100 met en œuvre un logiciel d'interception. Celui-ci comporte un module de duplication du type « port mirroring » pour dupliquer la totalité des requêtes HTTP transitant sur l'équipement 100.

25 Le logiciel d'interception comporte un module de filtrage permettant de filtrer les requêtes HTTP dupliquées comportant une URL faisant partie d'une liste d'URLs ou de parties d'URL de référence avec laquelle le module de filtrage est paramétré. L'URL du site Web surveillé est mentionnée dans cette liste de référence.

L'équipement d'interconnexion 100 est propre à router une requête HTTP interceptée vers l'un des serveurs de traitement 200, 202 du dispositif 150.

30 La figure 2 représente un programme dont l'exécution permet la mise en œuvre du procédé de traitement. Dans le mode de réalisation décrit, ce programme est réparti en différents logiciels, qui sont respectivement mémorisés et exécutés par des ordinateurs différents du dispositif 150.

Un logiciel de traitement 210 est mémorisé sur chacun des serveurs de traitement 200, 202.

35 Le logiciel de traitement 210 est capable de lire un fichier de configuration 211 contenant les différents paramètres nécessaires à son fonctionnement, tels que des

longueurs, exprimées en nombre de bits, correspondant aux longueurs des en-têtes (« HEADER ») des paquets des différentes couches OSI encapsulées dans une trame, des masques d'extraction de groupes de bits, et des valeurs prédéfinies attendues pour ces groupes de bits.

5 Le logiciel 210 comporte un module d'acquisition 212 capable d'écouter un port prédéfini du serveur de traitement, port sur lequel les trames interceptées sont incidentes. Le module 212 est capable d'acquérir l'intégralité d'une trame incidente sur le port écouté, de stocker la trame dans l'espace de mémoire adressable du serveur de traitement, et de placer, dans une pile 213 associée à une trame, un premier pointeur indiquant l'adresse
10 du premier bit de cette trame acquise.

 Le logiciel 210 comporte un module de sélection 214 apte à analyser en profondeur les trames acquises. Le module 214 est capable d'accéder, bit à bit, aux trames mémorisées dans l'espace de mémoire adressable du serveur de traitement. Le module de sélection est capable d'ajouter ou de soustraire des pointeurs de la pile 213
15 associée à une trame.

 Le module 214 comporte une pluralité de routines de vérification :

- une première routine de vérification d'une condition sur la couche IP, apte à comparer la longueur d'un paquet de bits inclus dans une trame avec une longueur prédéfinie de l'en-tête d'un paquet IP,
- 20 - une seconde routine de vérification d'une condition sur la couche IP, apte à appliquer un second masque adapté pour extraire un second groupe de bits, et à comparer ce second groupe de bit avec une seconde valeur binaire correspondant à une valeur attendue pour un paramètre de protocole présent dans un en-tête de paquet IP,
- une troisième routine de vérification d'une condition sur la couche TCP, apte à comparer
25 la longueur d'un paquet de bits inclus dans une trame avec une longueur prédéfinie de l'en-tête d'un paquet TCP,
- une quatrième routine de vérification d'une condition sur la couche HTTP, apte à appliquer un quatrième masque adapté pour extraire un quatrième groupe de bits, et à comparer ce quatrième groupe de bits avec une quatrième valeur binaire correspondant à
30 une valeur attendue pour un paramètre de type, présent dans un en-tête de paquet HTTP, et,
- une cinquième routine de vérification d'une condition sur la couche HTTP, apte à appliquer un cinquième masque adapté pour extraire un cinquième groupe de bits, et à comparer ce cinquième groupe de bits avec au moins une cinquième valeur binaire
35 correspondant à une valeur attendue pour au moins une portion d'un paramètre d'URL présent dans un en-tête de paquet HTTP.

Toutes ces vérifications sont effectuées sans désencapsuler les différentes couches du modèle OSI (IP, TCP et HTTP) permettant ainsi d'obtenir des temps de traitement réduits et donc de pouvoir analyser un flux de donnée ayant un débit très important.

5 Le logiciel 210 comporte également un module 216 d'extraction des données contenues dans un paquet HTTP. Le module 216 génère en sortie des données, et ajoute des meta-données associées. L'ensemble de ces données est appelé D.

10 Le logiciel de traitement 210 comporte un module 218 de sélection du serveur de stockage parmi les différents serveurs constituant l'étage de stockage 154. Le module 218 comporte une table d'occupation 219 donnant l'adresse des différents serveurs de stockage 300, 302, 304, ainsi que leur état respectif instantané d'occupation parmi des états « libre » et « occupé ».

15 Enfin, le logiciel de traitement 210 comporte un module 220 de codage et d'émission capable de prendre, en entrée, l'adresse du serveur choisi par le module 218, le port utilisé, et les données produites par le module 216, puis de communiquer ces données D au serveur de stockage choisi. Ces données peuvent être chiffrées en utilisant par exemple le code de cryptage AES 256 connu de l'homme du métier.

 Un logiciel de stockage 310 est exécuté sur chacun des serveurs de stockage 300, 302, 304.

20 Le logiciel de stockage 310 est capable de lire un fichier de configuration 311 contenant les différents paramètres nécessaires à son fonctionnement.

 Le logiciel 310 comporte un module d'acquisition 312 apte à écouter un port prédéfini du serveur de stockage et à acquérir les données D entrantes.

 Le logiciel 310 comporte un module de décodage 314 apte à extraire les données.

25 Le logiciel 310 comporte un module 316 propre à décoder les meta-données aux données D et à stocker l'ensemble de ces données dans un fichier F. Celui-ci est placé dans un répertoire particulier d'une structure d'archivage comportant une pluralité de répertoires.

30 Enfin, le logiciel 310 comporte un module de stockage 318 propre à surveiller le taux de remplissage de chacun des répertoires de la structure d'archivage, à comparer ce taux avec une valeur seuil, et à stocker le contenu d'un répertoire dans une table particulière de la base de données associée au serveur de stockage.

 Un logiciel de restitution 410 est propre à être exécuté par le serveur de restitution 400.

35 Le logiciel 410 comporte une interface homme/machine 112 permettant l'élaboration de requêtes complexes d'interrogation des bases de données 301, 303, 305.

Le logiciel 410 comporte un module 414 d'interrogation des bases de données. Il est capable d'interpréter une requête complexe en une pluralité de requêtes conformes au langage d'interrogation utilisé par les bases de données. Le module 414 est propre à émettre une requête d'interrogation vers les bases de données 301, 303, 305, et à recevoir les réponses correspondantes. Il est capable d'agréger ces réponses avant de les transmettre au module d'interface 412.

Le procédé d'analyse va maintenant être décrit en référence aux figures 3 et 4, la figure 5 rappelant la structure binaire d'une trame.

Le serveur 20 héberge un site Web sur lequel des utilisateurs échangent des données (telles que messages écrits, photos, vidéos, fichiers binaires), déposés sur le site et consultables au travers d'une page Web adaptée.

L'ALA souhaitant surveiller ce site Web met en œuvre le procédé pour acquérir des informations sur les utilisateurs de ce site Web.

L'ALA se rapproche alors du fournisseur d'accès à l'Internet gérant le premier réseau pour qu'il paramètre les différentes instances du logiciel d'interception avec, comme URL de référence, la racine du site Internet à surveiller. Les logiciels d'interception sont exécutés.

Lorsque l'utilisateur du poste client 10 dépose un message sur le site Web hébergé par le serveur 20, le poste client 10 émet une requête HTTP dont l'en-tête comporte la méthode « POST », de manière à ce que le serveur 20, destinataire, sache comment interpréter le message HTTP contenu dans la requête HTTP.

De manière similaire, lorsque l'utilisateur du poste 10 consulte une page du site Web, le poste client 10 émet une requête HTTP dont l'en-tête comporte la méthode « GET ».

Grâce au logiciel d'interception passive exécuté sur l'équipement d'interconnexion 100 les requêtes HTTP à destination du site Web accessible sur le serveur 20 et transitant sur l'équipement 100 sont interceptées. Elles sont dupliquées et leurs copies sont filtrées. Les requêtes HTTP comportant l'URL du site Web surveillé sont envoyées vers le dispositif 150. Les trames IP originales ne sont absolument pas impactées par le logiciel d'interception, ce qui garantit un fonctionnement normal du point de vue de l'utilisateur.

Le nombre des requêtes HTTP incidentes sur les serveurs de traitement est très élevé. La structure du dispositif 150 permet de répartir cette charge entre les différents serveurs de traitement.

Par l'exécution du logiciel de traitement 210, les étapes de traitement suivantes sont mises en œuvre au niveau du serveur 200.

Dans une étape initiale d'acquisition 612, le module 212 stocke une trame complète, correspondant à une requête HTTP incidente, dans l'espace mémoire adressable du serveur 200. Un premier pointeur P1 est placé dans une pile associée à cette trame. Le premier pointeur P1 indique l'adresse mémoire du premier bit de la trame à filtrer.

Puis, le procédé se poursuit par une étape de sélection 614 consistant en une analyse, en profondeur, de la structure binaire de la trame.

Comme cela est représenté en détail sur la figure 4, l'étape de sélection 614, qui est réalisée par l'exécution du module de sélection 214, débute par la détermination de la longueur L0 de la trame (étape 1010 sur la figure 4).

L'en-tête de la couche de transport d'une trame (couches 2 du modèle OSI) ayant une première longueur prédéterminée L1, un second pointeur P2 est placé dans la pile associée à la trame. Ce second pointeur pointe vers une adresse de l'espace mémoire obtenue par décalage d'une longueur L1 de l'adresse indiquée par le premier pointeur P1 (étape 1020). Ainsi, le second pointeur pointe sur le premier octet de la couche IP de la trame (couche de niveau 3 du modèle OSI).

La longueur L2 du paquet IP encapsulé dans la trame est calculée à l'étape 1030. Cette longueur L2 est obtenue en soustrayant à la longueur L0, la longueur L1.

La longueur L3 de l'en-tête d'un paquet IP est définie par le protocole IP. Cette longueur L3 permet de vérifier une première condition qui consiste à comparer la longueur L2 du paquet IP à la longueur L3 (étape 1040).

Si la longueur L2 est inférieure à la longueur L3, cela signifie que le paquet considéré n'est pas un paquet IP. En conséquence, la trame est rejetée et le procédé passe à la sélection de la trame suivante.

En revanche, si la longueur L2 est supérieure à la longueur L3, cela signifie que, s'il s'agit effectivement d'un paquet IP, celui-ci comporte, en plus d'un en-tête IP, un message IP comportant éventuellement des données pertinentes.

A l'étape 1050, un second masque M2 est appliqué sur l'en-tête IP du paquet IP (« HEADER » du paquet IP) de manière à extraire un second groupe de bits et à le comparer à une seconde valeur binaire attendue d'un second paramètre relatif au protocole utilisé dans la couche de transport (couche de niveau 4 du modèle OSI), second paramètre présent dans l'en-tête IP. Dans le présent mode de réalisation, la seconde valeur attendue correspond à l'utilisation du protocole TCP.

A l'issue de la vérification de cette seconde condition, si la valeur du second paramètre de protocole est différente de « TCP », la trame est rejetée et le procédé passe à la sélection de la trame suivante.

En revanche, si la valeur du second paramètre de protocole est égale à « TCP », un troisième pointeur P3 est placé, à l'étape 1060, dans la pile 213 associée à la trame. Ce troisième pointeur pointe sur une adresse obtenue par décalage d'une longueur L3 de l'adresse indiquée par le second pointeur P2. Le troisième pointeur indique le début de la
5 couche TCP de la trame.

A l'étape 1070, une longueur L4 est calculée qui correspond à la longueur du paquet TCP. Cette longueur L4 est obtenue par la différence entre la longueur L2 et la longueur L3.

La longueur L5 de l'en-tête d'un paquet TCP est prédéterminée. Cette longueur L5 permet la tester une troisième condition qui consiste à comparer la longueur L4 du paquet TCP à la longueur L5 (étape 1080).
10

Si la longueur L4 est inférieure à la longueur L5, cela signifie que le paquet considéré n'est pas un paquet TCP. En conséquence, la trame est rejetée et le procédé passe à la sélection de la trame suivante.

En revanche, si la longueur L4 est supérieure à la longueur L5, le paquet TCP comporte, en plus d'un en-tête TCP, un message TCP pouvant contenir des informations pertinentes.
15

A l'étape 1090, un quatrième pointeur P4 est placé dans la pile associée à la trame. Ce quatrième pointeur pointe vers une adresse qui correspond au décalage d'une longueur L5 de l'adresse indiqué par le troisième pointeur P3. Le quatrième pointeur pointe sur le début de la couche HTTP de la trame étudiée (couches d'application 5 à 7 du modèle OSI).
20

Puis, à l'étape 1100, un quatrième masque M4 est appliqué sur l'en-tête HTTP de manière à extraire un quatrième groupe de bits et à le comparer à une quatrième valeur binaire attendue pour un quatrième paramètre de type du paquet HTTP. La quatrième valeur attendue est la valeur « POST » ou la valeur « GET » de ce paramètre de méthode.
25

Si la méthode HTTP utilisée n'est pas l'une des deux méthodes précédentes, la trame n'est pas considérée et le procédé passe à l'étape de sélection de la trame suivante.
30

Si la méthode HTTP est un POST ou un GET, à l'étape 1110, un cinquième masque M5 est appliqué sur l'en-tête HTTP de manière à comparer une partie de l'URL à une pluralité de cinquième valeurs non-souhaitées correspondant à des chaînes de caractères de référence.

Si la comparaison s'avère positive, la trame est rejetée sinon la trame est sélectionnée.
35

Ce dernier test permet, par exemple, d'écarter les requêtes HTTP comportant un message correspondant à une image, en mentionnant la chaîne « .jpg » dans la liste des chaînes de caractères de référence.

Pour une trame sélectionnée, le procédé se poursuit par une étape 616 d'extraction et de reformatage des données HTTP par exécution du module 216. Les données extraites de l'en-tête HTTP de la requête HTTP sont l'URL, l'adresse IP source de la trame, l'adresse IP destinataire de la trame, le « User-Agent » c'est-à-dire l'identifiant du navigateur utilisé, et le « REFERER » c'est-à-dire l'URL d'une page Web sur laquelle se situe un lien du type HyperText, que le client vient de suivre, pour accéder à une ressource du site Web surveillé. Cela peut être un lien sur une page extérieure au site Web surveillé, mais également un lien sur le site Web surveillé.

Chacune de ces données est conservée dans une variable associée.

Avantageusement, des données supplémentaires, dites meta-données, sont associées à la trame traitée. Ainsi, si l'URL de la requête HTTP correspond à une URL0 de référence qui, dans le fichier de configuration 211, est associée à un type d'affaire particulier, tel que le type « terrorisme », le type d'affaire constitue une meta-donnée associée à la trame lors de l'étape 616.

Un ensemble de données et de meta-données, composant un message de données D, est finalement mémorisé dans un espace mémoire tampon du serveur de traitement 200.

A l'étape 618, le module de sélection 218 surveillant cet espace mémoire tampon, constate qu'un nouveau message de données vient d'y être déposé afin d'être envoyé vers une base de données de stockage.

Le module 218 lit la table 219 à la recherche de l'adresse d'un serveur de stockage 300, 302, 304 dans l'état « libre » auquel adresser le message de données. Le module 218 sélectionne un serveur de stockage destinataire, par exemple le serveur de stockage 300.

Le message de données est donc envoyé au serveur de stockage choisi. Ce message peut être chiffré en AES 256. Du côté du serveur de stockage 300, après une étape 712 d'acquisition du message de données D, une étape de décodage 714 permet de récupérer les données D qui sont stockées dans un fichier F.

Une étape de classification 716 du fichier de données permet alors de choisir un répertoire d'archivage pour ce fichier. Le choix d'un répertoire particulier s'effectue sur la base des meta-données qui ont été associées au fichier F.

L'étape de stockage dans une base de données 301 associée au serveur de stockage 300, étape 718 sur la figure 3, est réalisée par l'exécution du module 318 qui

scrute en permanence le taux de remplissage de chacun des répertoires de la structure d'archivage. Lorsque le taux de remplissage d'un répertoire dépasse un taux seuil prédéterminé, l'intégralité du contenu de ce répertoire est sauvegardée dans la base de données 301, dans un tableau ayant un format prédéterminé.

5 A l'étape 812, hors ligne, au travers de l'interface homme/machine 412 affichée sur l'écran du serveur de restitution 400, un membre de l'ALA construit des requêtes complexes d'interrogation des bases de données 301, 303, 305. ce membre utilise un meta-langage.

10 A l'étape 814, ces requêtes complexes sont transmises au module de consultation 414 qui les traduit en autant de requêtes utilisant le langage SQL permettant une interrogation directe des bases de données 301, 303 et/ou 305. Les données extraites des différentes bases de données sont rapatriées sur le serveur de restitution 400. Le module de consultation 414 agrège ces différentes données pour qu'elles soient présentées à l'opérateur à travers l'interface 412.

15 Le dispositif et le procédé de traitement qui viennent d'être décrits permettent de traiter un flux de données de grand volume par un unique ordinateur serveur de traitement comportant une carte mère présentant des caractéristiques standard. L'échelle du dispositif de traitement étant facilement adaptable aux besoins, en multipliant les ordinateurs composant chacune des couches du dispositif, des flux de données très
20 élevés peuvent être traités par le dispositif selon l'invention. Ces flux de données élevés sont typiquement ceux que l'on trouve au point d'accès d'un sous-réseau national du réseau Internet.

25 Par le traitement en profondeur d'une requête HTTP, c'est-à-dire au niveau binaire de la trame correspondante, le procédé évite une multiplication des temps de calcul et un allongement considérable des temps de traitement requis pour chaque requête, tout en permettant l'extraction d'un grand nombre de données nécessaires à la surveillance du site Web et des activités de ses utilisateurs.

REVENDEICATIONS

1.- Procédé d'analyse de requête HTTP interceptées sur un réseau IP (30) pour la surveillance de l'activité des utilisateurs d'un site Web prédéterminé, comportant les étapes consistant à :

- acquérir (612) la trame de données complète d'une requête HTTP;
- sélectionner (614) la trame de données acquise si sa structure binaire vérifie une pluralité de conditions comportant au moins une condition correspondant à la couche IP (1040, 1050) de la trame, au moins une condition correspondant à la couche de transport (1080) de la trame et au moins une condition correspondant à la couche d'application (1100, 1110) de la trame;
- extraire (616) des données d'intérêt de la couche d'application de la trame sélectionnée ;
- enregistrer (618, 620, 712, 714, 716, 718) les données extraites, dans une base de données (301, 303, 305).

2.- Procédé selon la revendication 1, caractérisé en ce que l'étape de sélection (612) permet la sélection d'une trame dont la couche de transport est une couche TCP et dont la couche d'application est une couche HTTP.

3.- Procédé selon la revendication 2, caractérisé en ce que, à l'étape de sélection (612), ladite au moins une condition sur la couche IP, respectivement ladite au moins une condition sur la couche TCP, consiste à comparer (1040 ; 1080) la longueur d'un paquet de bits inclus dans la trame acquise, paquet considéré comme étant un paquet IP, respectivement un paquet TCP, avec une longueur prédéfinie de l'en-tête d'un paquet IP, respectivement d'un paquet TCP.

4.- Procédé selon la revendication 2 ou la revendication 3, caractérisé en ce que, à l'étape de sélection (612), ladite au moins une condition sur la couche IP, respectivement ladite au moins une condition sur la couche HTTP, consiste à appliquer, sur l'en-tête d'un paquet de bits inclus dans la trame acquise, paquet considéré comme étant un paquet IP, respectivement un paquet HTTP, un masque pour extraire un groupe de bits et à comparer (1050 ; 1100, 1110) ce groupe de bits avec valeur binaire attendue pour un paramètre présent dans l'en-tête d'un paquet IP, respectivement l'en-tête d'un paquet HTTP.

5.- Procédé selon l'une quelconque des revendications précédentes, caractérisé en ce que, entre l'étape consistant à extraire les données de la couche d'application de ladite trame et à enregistrer ces données dans une base de données, le procédé comporte une étape supplémentaire consistant à mettre en forme les données extraites selon un modèle prédéterminé, de préférence en y associant des meta-données.

6.- Dispositif pour la mise en œuvre du procédé selon l'une quelconque des revendications 1 à 5, caractérisé en ce qu'il comporte :

- un moyen d'acquisition (212) d'une trame de données complète d'une requête HTTP interceptée sur un réseau de communication IP auquel ledit dispositif est connecté ;
- un moyen de sélection (214) propre à vérifier une pluralité de conditions sur la structure binaire d'une trame de données acquise obtenue en sortie du moyen d'acquisition, et comportant au moins une routine de vérification d'une condition correspondant à la couche IP de la trame, au moins une routine de vérification d'une condition correspondant à la couche de transport de la trame et au moins une routine de vérification d'une condition correspondant à la couche d'application de la trame ;
- un moyen d'extraction (216) apte à extraire des données de la couche d'application d'une trame de données sélectionnée obtenue en sortie du moyen de sélection ;
- un moyen d'enregistrement (218 à 318) apte à stocker les données extraites obtenues en sortie du module d'extraction dans une base de données (301, 303, 305).

7.- Dispositif selon la revendication 6, caractérisé en ce que le moyen de sélection (212) est adapté pour la sélection d'une trame de données acquises dont la couche de transport est une couche TCP et dont la couche d'application est une couche HTTP.

8.- Dispositif selon la revendication 6 ou la revendication 7, caractérisé en ce qu'il comporte un étage de traitement (152) comportant une pluralité d'ordinateurs serveur de traitement (200, 202), chaque ordinateur serveur de traitement étant connecté audit réseau de communication IP et comportant une instanciation desdits moyens d'acquisition, de sélection et d'extraction.

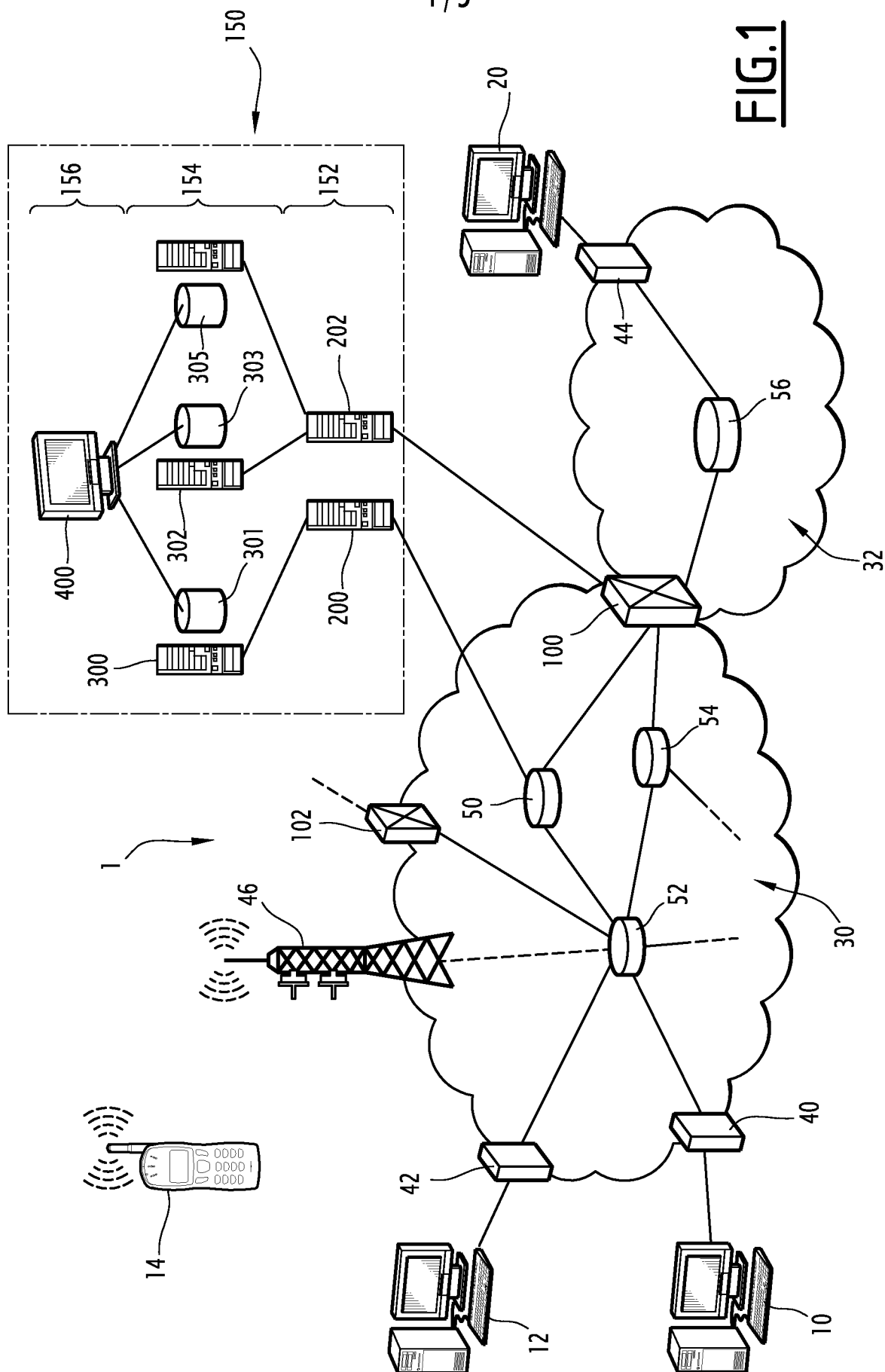
9.- Dispositif selon la revendication 8, caractérisé en ce qu'il comporte, en outre, un étage de stockage (154) comportant une pluralité d'ordinateurs serveur de stockage (300, 3020, 304), chaque ordinateur serveur de stockage étant connecté à ladite pluralité d'ordinateurs serveur de traitement, étant associé à au moins une base de données (301, 303, 305), et comportant une instanciation desdits moyens de stockage apte à stocker les données extraites communiquées par un ordinateur serveur de traitement dans la base de données associée à l'ordinateur serveur de stockage considéré.

15

10.- Dispositif selon la revendication 9, caractérisé en ce qu'il comporte, en outre, un étage de restitution (156) comportant au moins un ordinateur de restitution (400) comportant des moyens d'interrogation des différentes bases de données de l'étage de stockage.

1/5

FIG.1



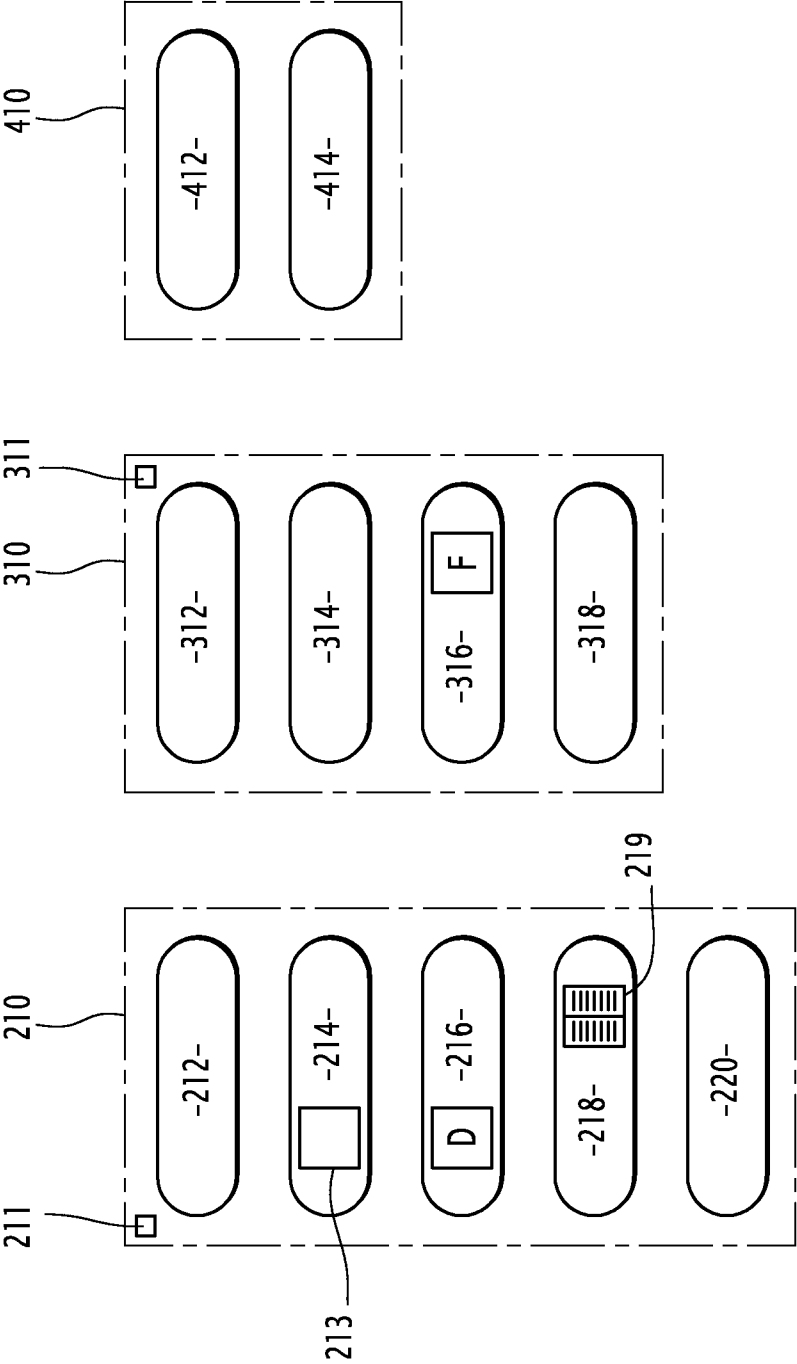


FIG.2

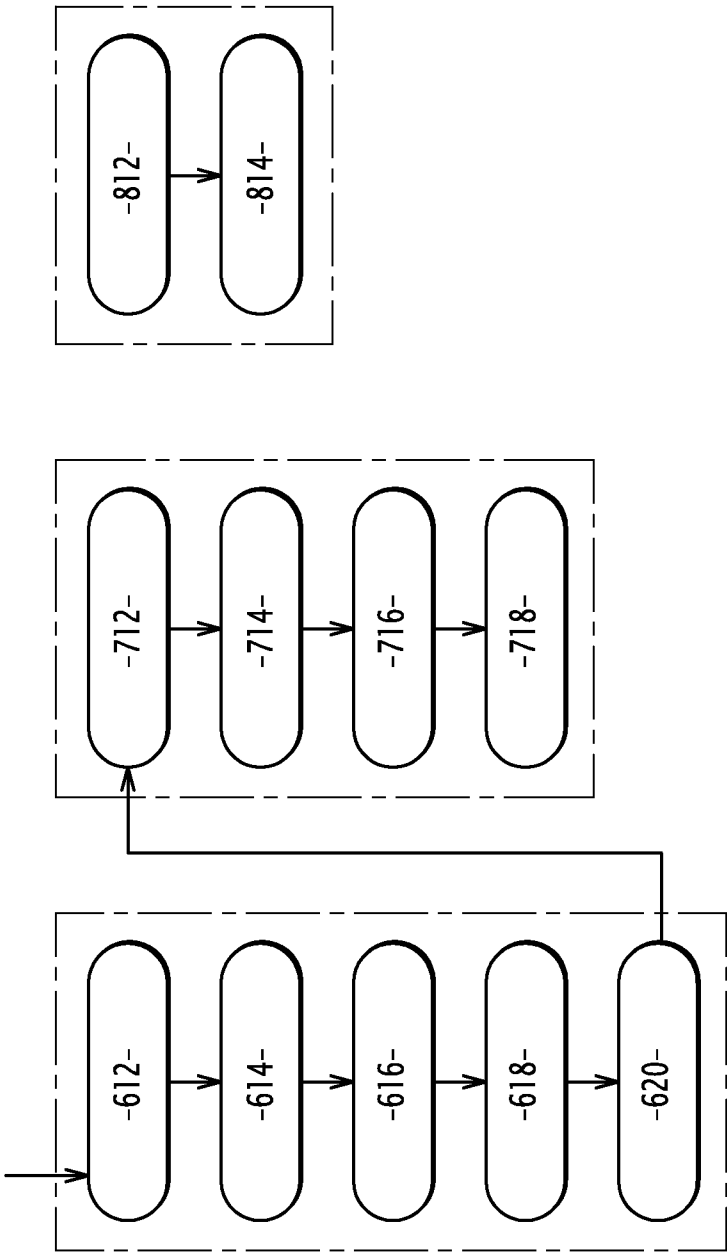
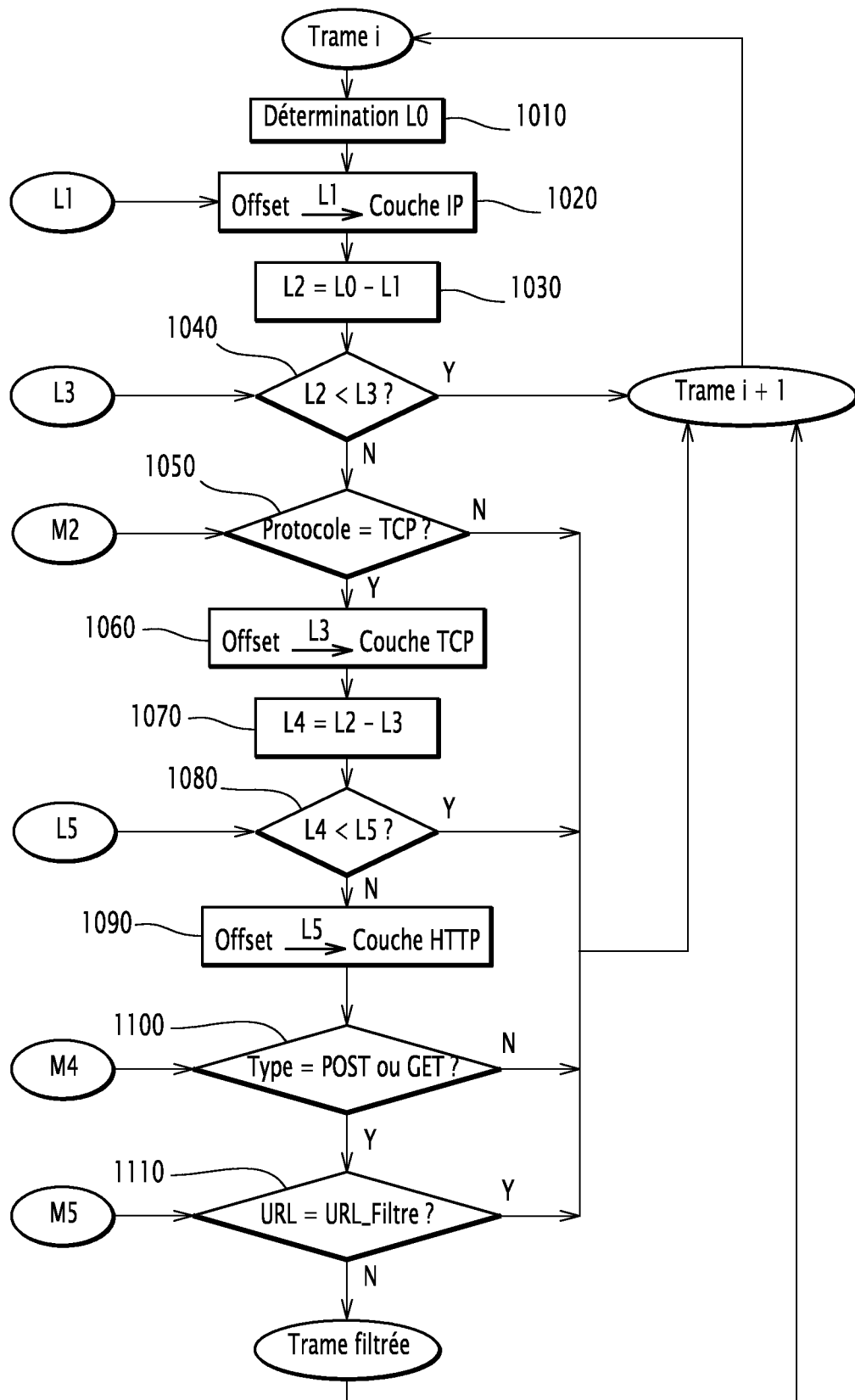
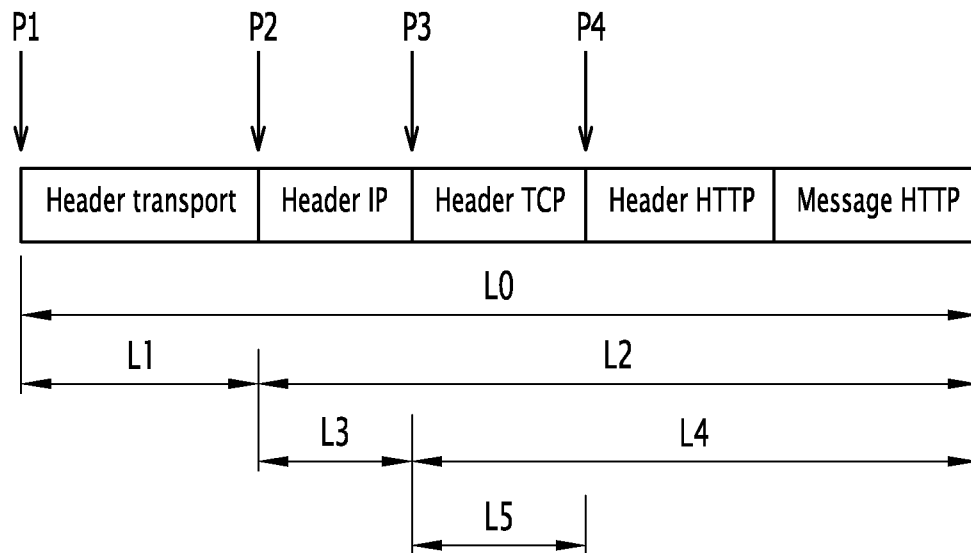


FIG.3

4/5

FIG.4

5/5

FIG.5

INTERNATIONAL SEARCH REPORT

International application No
PCT/FR2011/051153

A. CLASSIFICATION OF SUBJECT MATTER

INV. H04L29/06 H04L29/08 H04L12/26
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
H04L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

EPO-Internal, WPI Data, COMPENDEX, INSPEC

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	MAO ET. AL: "Cluster-based Online Monitoring System of Web Traffic", ACM, 2 PENN PLAZA, SUITE 701 - NEW YORK USA, 9 November 2001 (2001-11-09), pages 47-53, XP040131337, figures 1-2 abstract paragraph [0001] - paragraph [03.1] paragraph [03.3] - paragraph [03.4] paragraph [0006]	1-10
X	WO 2005/077009 A2 (NARUS INC [US]; CHANDRA ROHIT [US]) 25 August 2005 (2005-08-25) figures 2-7 paragraph [0001] paragraphs [0025] - [0044] ----- -/-	1-10



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier document but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.
"&" document member of the same patent family

Date of the actual completion of the international search

19 July 2011

Date of mailing of the international search report

27/07/2011

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Jeampierre, Gérald

INTERNATIONAL SEARCH REPORT

International application No

PCT/FR2011/051153

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	AHMED M K ET AL: "An Automated User Transparent Approach to log Web URLs for Forensic Analysis", IT SECURITY INCIDENT MANAGEMENT AND IT FORENSICS, 2009. IMF '09. FIFTH INTERNATIONAL CONFERENCE ON, IEEE, PISCATAWAY, NJ, USA, 15 September 2009 (2009-09-15), pages 120-127, XP031540596, ISBN: 978-0-7695-3807-5 figures 1-2 table 2 abstract paragraph [0001] - paragraph [0002] paragraph [0003] - paragraph [03.1] -----	1-10
A	JP 2004 145583 A (NIPPON TELEGRAPH & TELEPHONE) 20 May 2004 (2004-05-20) Une traduction automatique par ordinateur de cette publication japonaise peut être trouvée sur le site web de l'Office Japonais des Brevets: http://www.ipdl.ncipi.go.jp/homepg_e.ipdl. figure 12 paragraphs [0002] - [0003] paragraph [0006] paragraph [0012] paragraph [0014] paragraphs [0023] - [0024] -----	1-10

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/FR2011/051153

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2005077009 A2	25-08-2005	US 2005198268 A1	08-09-2005
JP 2004145583 A	20-05-2004	NONE	

RAPPORT DE RECHERCHE INTERNATIONALE

Demande internationale n°

PCT/FR2011/051153

A. CLASSEMENT DE L'OBJET DE LA DEMANDE INV. H04L29/06 H04L29/08 H04L12/26 ADD.		
Selon la classification internationale des brevets (CIB) ou à la fois selon la classification nationale et la CIB		
B. DOMAINES SUR LESQUELS LA RECHERCHE A PORTE Documentation minimale consultée (système de classification suivi des symboles de classement) H04L		
Documentation consultée autre que la documentation minimale dans la mesure où ces documents relèvent des domaines sur lesquels a porté la recherche		
Base de données électronique consultée au cours de la recherche internationale (nom de la base de données, et si cela est réalisable, termes de recherche utilisés) EPO-Internal, WPI Data, COMPENDEX, INSPEC		
C. DOCUMENTS CONSIDERES COMME PERTINENTS		
Catégorie*	Identification des documents cités, avec, le cas échéant, l'indication des passages pertinents	no. des revendications visées
X	MAO ET. AL: "Cluster-based Online Monitoring System of Web Traffic", ACM, 2 PENN PLAZA, SUITE 701 - NEW YORK USA, 9 novembre 2001 (2001-11-09), pages 47-53, XP040131337, figures 1-2 abrégé alinéa [0001] - alinéa [03.1] alinéa [03.3] - alinéa [03.4] alinéa [0006]	1-10
X	WO 2005/077009 A2 (NARUS INC [US]; CHANDRA ROHIT [US]) 25 août 2005 (2005-08-25) figures 2-7 alinéa [0001] alinéas [0025] - [0044]	1-10
----- -/-		
☒ Voir la suite du cadre C pour la fin de la liste des documents ☒ Les documents de familles de brevets sont indiqués en annexe		
* Catégories spéciales de documents cités:		
"A" document définissant l'état général de la technique, non considéré comme particulièrement pertinent "E" document antérieur, mais publié à la date de dépôt international ou après cette date "L" document pouvant jeter un doute sur une revendication de priorité ou cité pour déterminer la date de publication d'une autre citation ou pour une raison spéciale (telle qu'indiquée) "O" document se référant à une divulgation orale, à un usage, à une exposition ou tous autres moyens "P" document publié avant la date de dépôt international, mais postérieurement à la date de priorité revendiquée	"T" document ultérieur publié après la date de dépôt international ou la date de priorité et n'appartenant pas à l'état de la technique pertinent, mais cité pour comprendre le principe ou la théorie constituant la base de l'invention "X" document particulièrement pertinent; l'invention revendiquée ne peut être considérée comme nouvelle ou comme impliquant une activité inventive par rapport au document considéré isolément "Y" document particulièrement pertinent; l'invention revendiquée ne peut être considérée comme impliquant une activité inventive lorsque le document est associé à un ou plusieurs autres documents de même nature, cette combinaison étant évidente pour une personne du métier "&" document qui fait partie de la même famille de brevets	
Date à laquelle la recherche internationale a été effectivement achevée	Date d'expédition du présent rapport de recherche internationale	
19 juillet 2011	27/07/2011	
Nom et adresse postale de l'administration chargée de la recherche internationale Office Européen des Brevets, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016		Fonctionnaire autorisé Jeampierre, Gérald

C(suite). DOCUMENTS CONSIDERES COMME PERTINENTS		
Catégorie*	Identification des documents cités, avec, le cas échéant, l'indication des passages pertinents	no. des revendications visées
X	AHMED M K ET AL: "An Automated User Transparent Approach to log Web URLs for Forensic Analysis", IT SECURITY INCIDENT MANAGEMENT AND IT FORENSICS, 2009. IMF '09. FIFTH INTERNATIONAL CONFERENCE ON, IEEE, PISCATAWAY, NJ, USA, 15 septembre 2009 (2009-09-15), pages 120-127, XP031540596, ISBN: 978-0-7695-3807-5 figures 1-2 tableau 2 abrégé alinéa [0001] - alinéa [0002] alinéa [0003] - alinéa [03.1] -----	1-10
A	JP 2004 145583 A (NIPPON TELEGRAPH & TELEPHONE) 20 mai 2004 (2004-05-20) Une traduction automatique par ordinateur de cette publication japonaise peut être trouvée sur le site web de l'Office Japonais des Brevets: http://www.ipdl.ncipi.go.jp/homepg_e.ipdl. figure 12 alinéas [0002] - [0003] alinéa [0006] alinéa [0012] alinéa [0014] alinéas [0023] - [0024] -----	1-10

RAPPORT DE RECHERCHE INTERNATIONALE

Renseignements relatifs aux membres de familles de brevets

Demande internationale n°

PCT/FR2011/051153

Document brevet cité au rapport de recherche	Date de publication	Membre(s) de la famille de brevet(s)	Date de publication
WO 2005077009 A2	25-08-2005	US 2005198268 A1	08-09-2005
JP 2004145583 A	20-05-2004	AUCUN	