US 20140304653A1

(19) United States

(12) Patent Application Publication

(10) Pub. No.: US 2014/0304653 A1

(43) Pub. Date: Oct. 9, 2014

Fischer-Wuenschel et al.

(54) METHOD FOR GENERATING RULES AND PARAMETERS FOR ASSESSING RELEVANCE OF INFORMATION DERIVED FROM INTERNET TRAFFIC

(71) Applicant: **GfK US Holdings, Inc.**, New York, NY (US)

(72) Inventors: **Bernhard Fischer-Wuenschel**, Weihenzell (DE); **Thomas Ruf**, Fuerth (DE); **Renate Wendlik**, Roth (DE)

(73) Assignee: **GfK US Holdings, Inc.**, New York, NY (US)

(21) Appl. No.: **14/310,817**

(22) Filed: **Jun. 20, 2014**

Related U.S. Application Data

(63) Continuation of application No. 13/157,106, filed on Jun. 9, 2011, now Pat. No. 8,818,927.

Publication Classification

(51) **Int. Cl.**

G06F 3/0481	(2006.01)
G06F 3/0484	(2006.01)
G06F 3/0482	(2006.01)

(52) **U.S. Cl.**

CPC ***G06F 3/04812*** (2013.01); ***G06F 3/0482*** (2013.01); ***G06F 3/04842*** (2013.01)

USPC **715/810**

(57) **ABSTRACT**

A method is disclosed for generating a set of optimal rules and parameters for use by an evaluation engine on a volume of information extracted from a stream of IP packets associated with a web browsing session conducted over a network in order to filter user-initiated traffic flowing across the network from non-user-initiated traffic. Deep packet inspection is performed to extract the volume of information from the stream that conforms to at least one discrimination criteria. An initial iteration of application of the evaluation engine to the volume is performed by selecting initial discrimination criteria and rules for generating the filtered results and a distance between the filtered results and known actual user-initiated traffic is measured. Subsequent iterations of application of the evaluation engine to the volume are performed by changing the discrimination criteria and/or rules until there is no significant improvement in the measured distance.

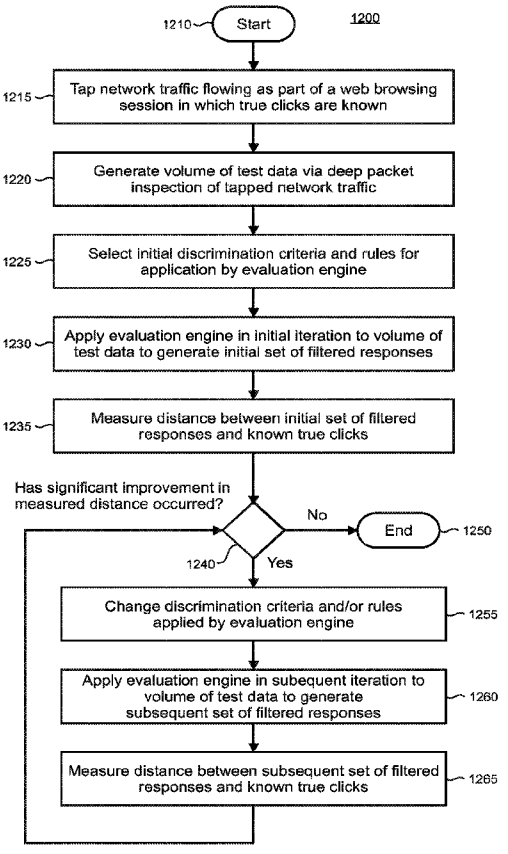


FIG. 1

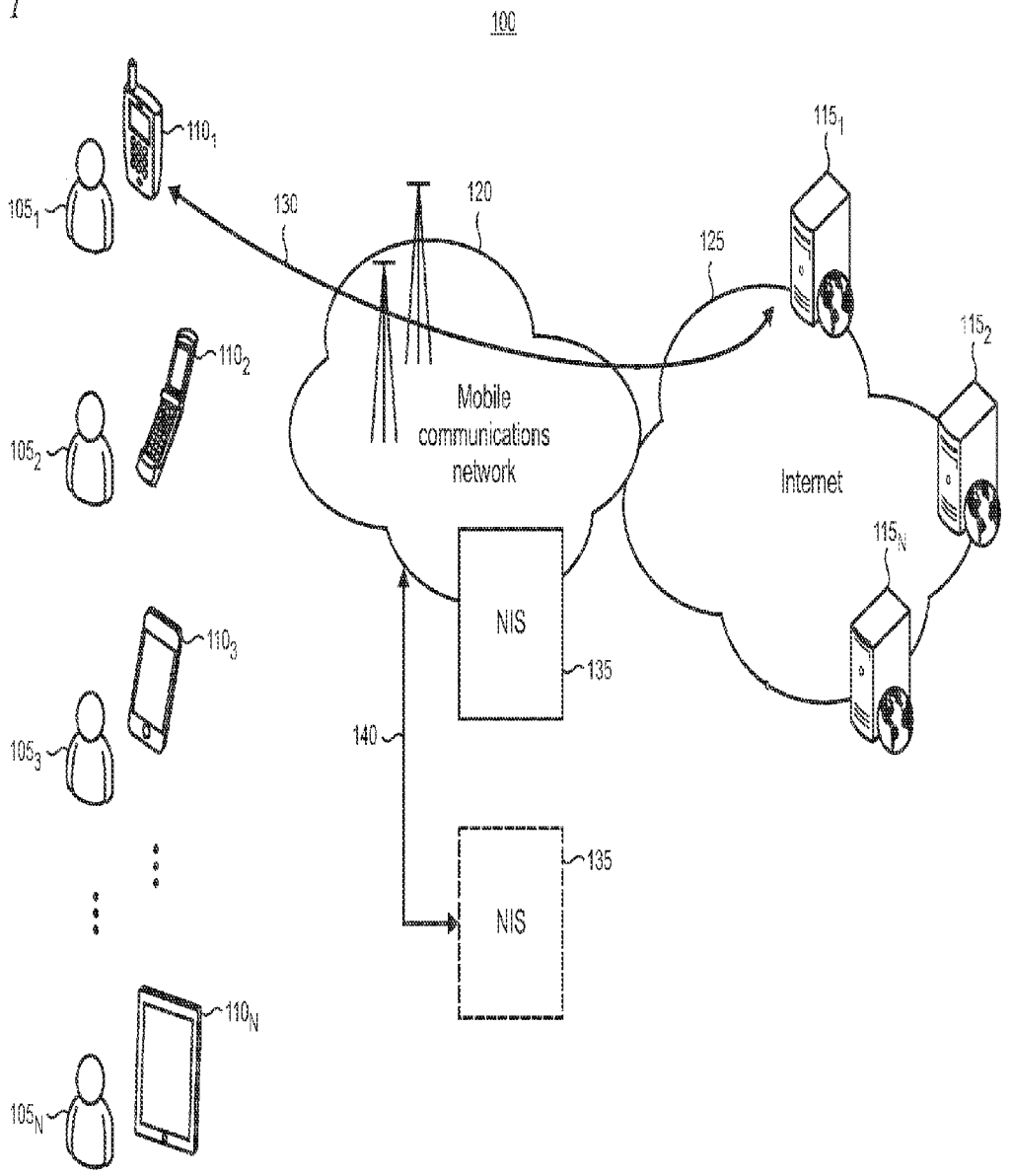

100
110_1
105_1
130
120
115_1
125
115_2
110_2
105_2
Mobile communications network
Internet
115_N
NIS
135
140
110_3
105_3
135
NIS
110_N
105_N

FIG. 2

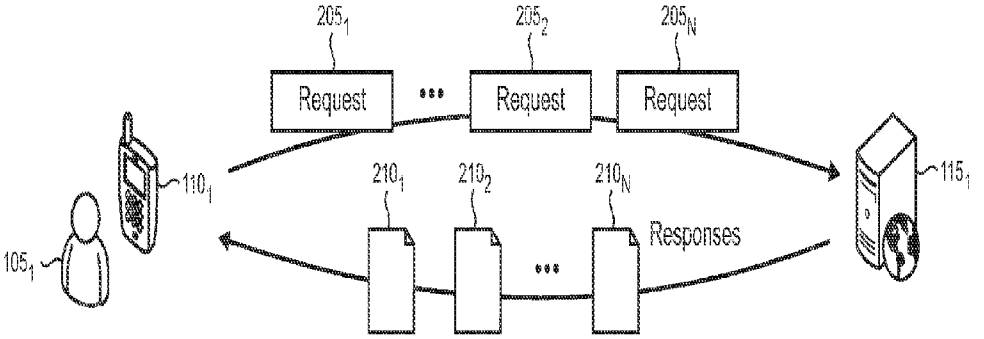

205_1
205_2
205_N
Request
Request
Request
110_1
115_1
105_1
210_1
210_2
210_N
Responses

FIG. 3

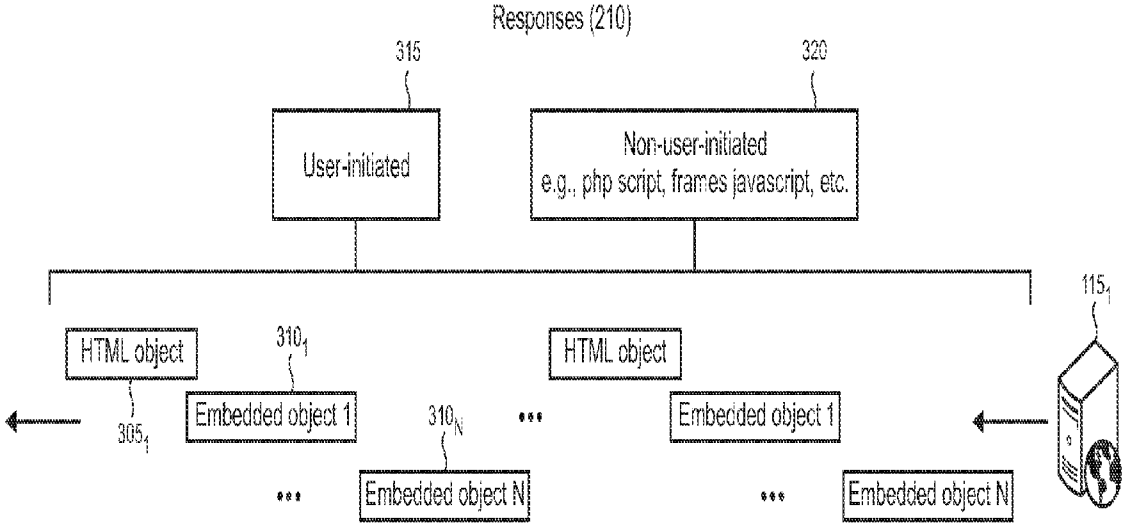

Responses (210)
315
320
User-initiated
Non-user-initiated
e.g., php script, frames javascript, etc.
115_1
HTML object
310_1
HTML object
305_1
Embedded object 1
310_N
Embedded object 1
Embedded object N
Embedded object N

FIG. 4

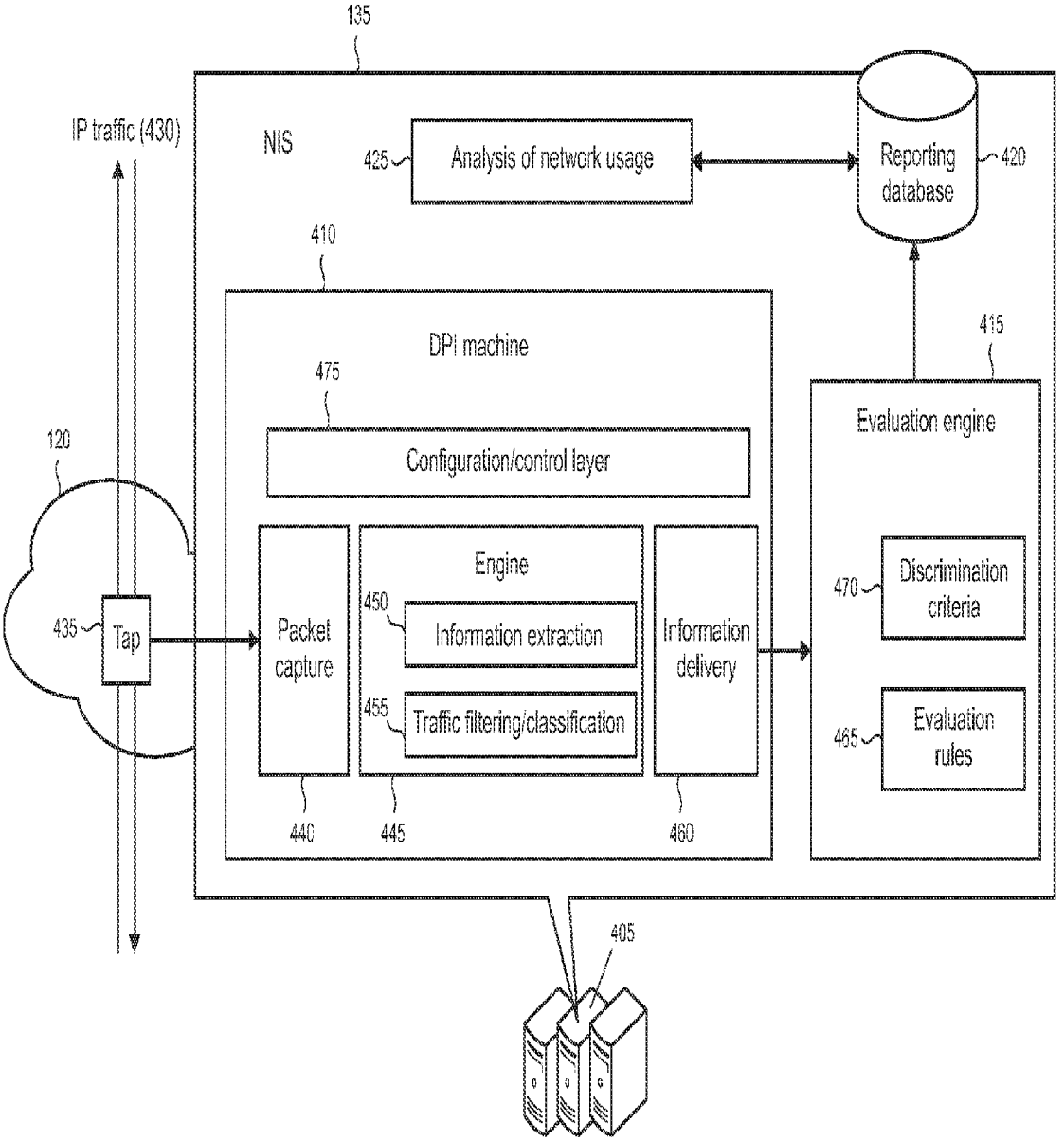
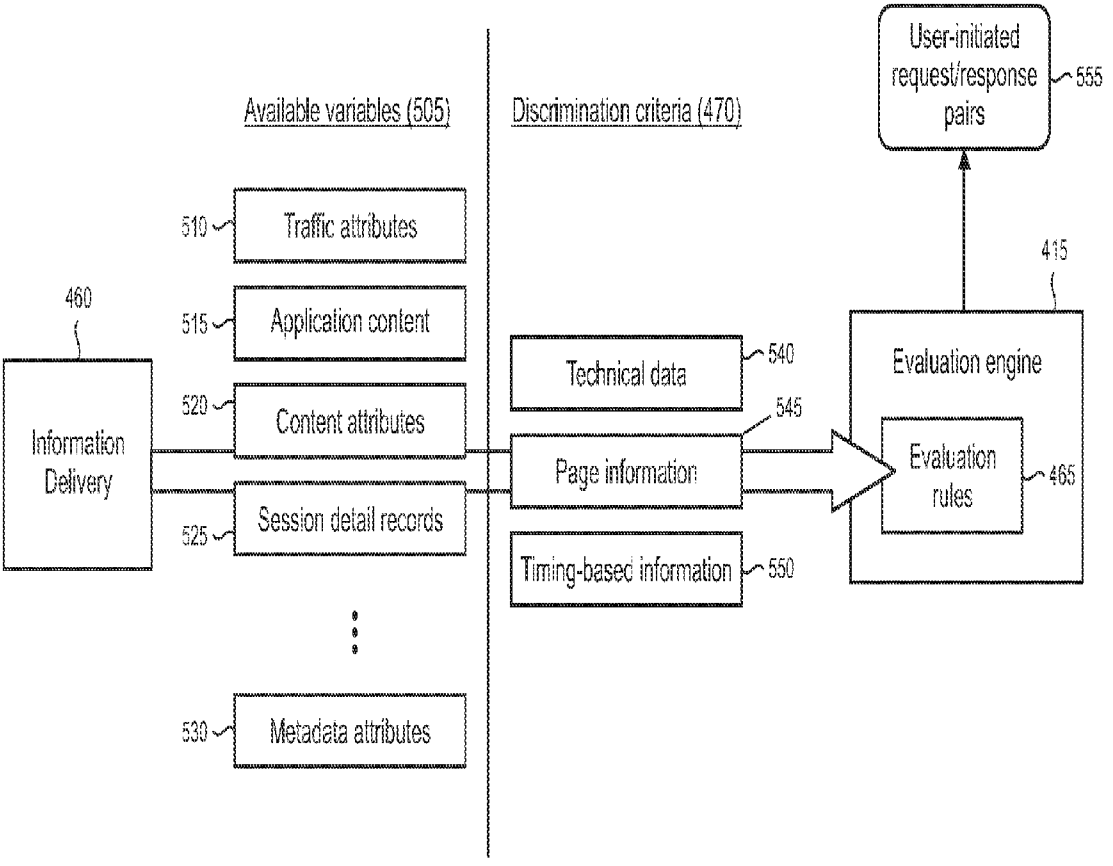

135
IP traffic (430)
NIS
425
Analysis of network usage
Reporting database
420
410
DPI machine
475
Configuration/control layer
415
Evaluation engine
120
435
Tap
Packet capture
Engine
450
Information extraction
455
Traffic filtering/classification
Information delivery
470
Discrimination criteria
465
Evaluation rules
440
445
460
405

FIG. 5


User-initiated request/response pairs
555
Available variables (505)
Discrimination criteria (470)
510
Traffic attributes
415
460
515
Application content
Technical data
540
Evaluation engine
520
Content attributes
545
Information Delivery
Page information
Evaluation rules
465
525
Session detail records
Timing-based information
550
530
Metadata attributes

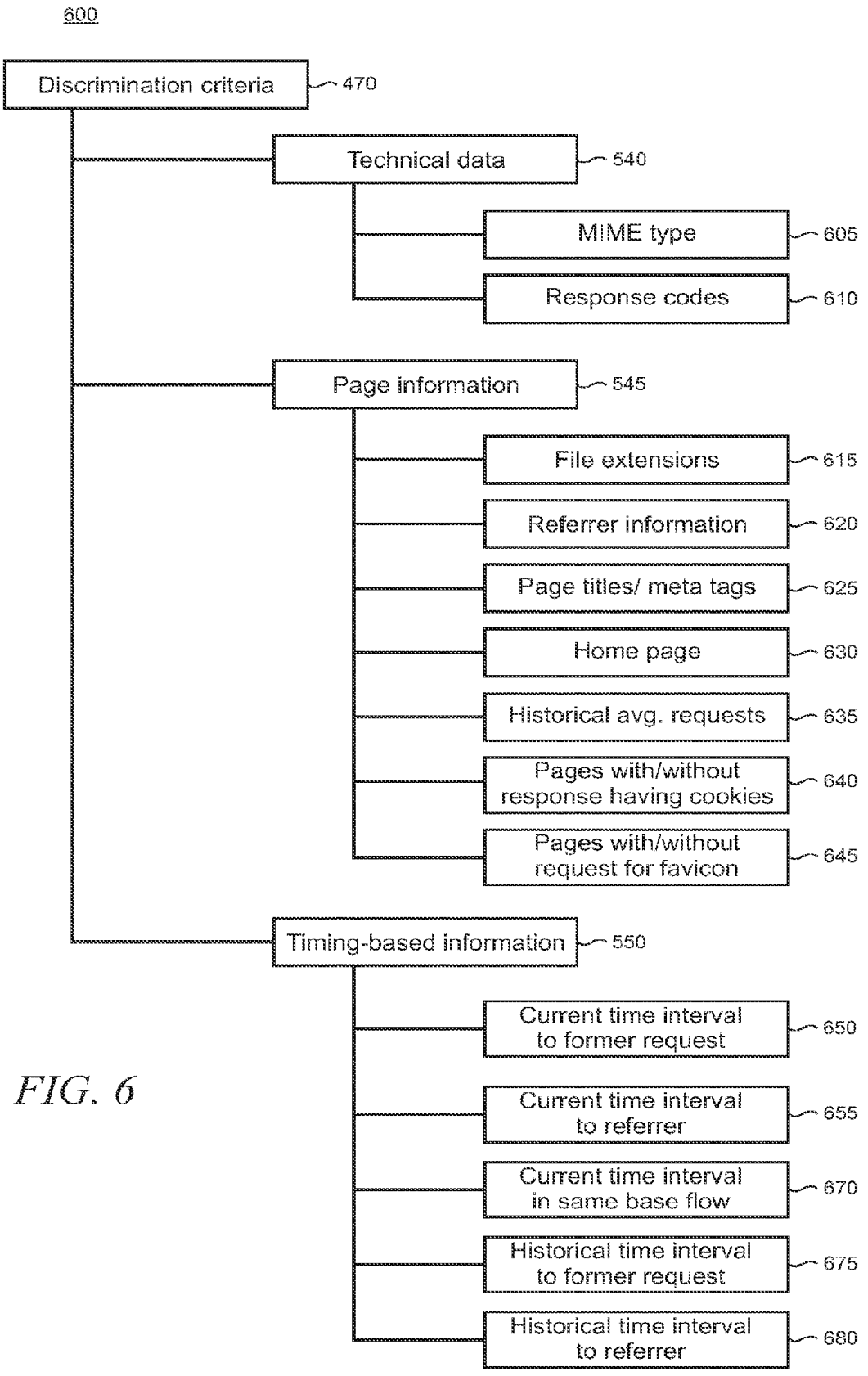

600
Discrimination criteria
470
Technical data
540
MIME type
605
Response codes
610
Page information
545
File extensions
615
Referrer information
620
Page titles/ meta tags
625
Home page
630
Historical avg. requests
635
Pages with/without response having cookies
640
Pages with/without request for favicon
645
Timing-based information
550
Current time interval to former request
650
Current time interval to referrer
655
Current time interval in same base flow
670
Historical time interval to former request
675
Historical time interval to referrer
680

FIG. 6

FIG. 7

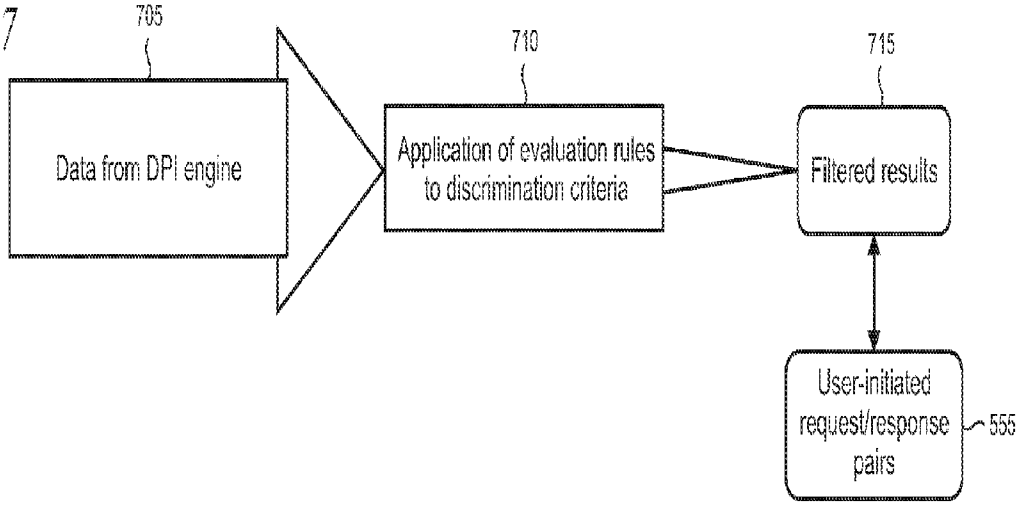

705
Data from DPI engine
710
Application of evaluation rules to discrimination criteria
715
Filtered results
User-initiated request/response pairs
555

FIG. 8

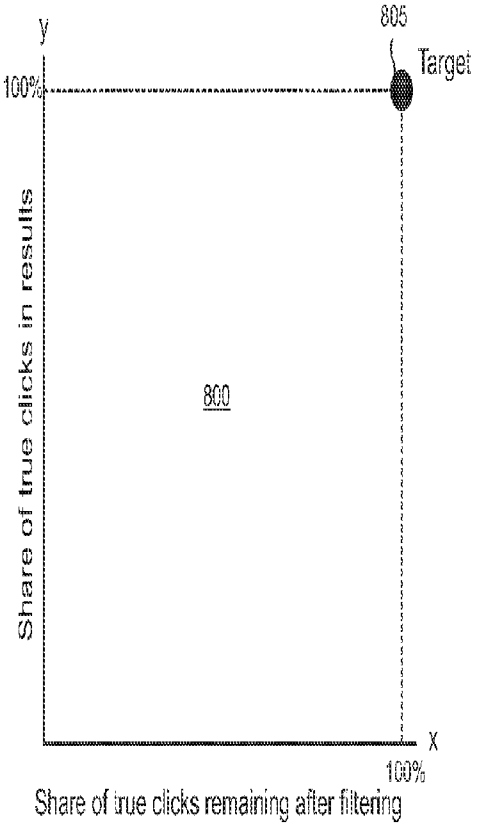

y
805
Target
100%
Share of true clicks in results
800
X
100%
Share of true clicks remaining after filtering

FIG. 9

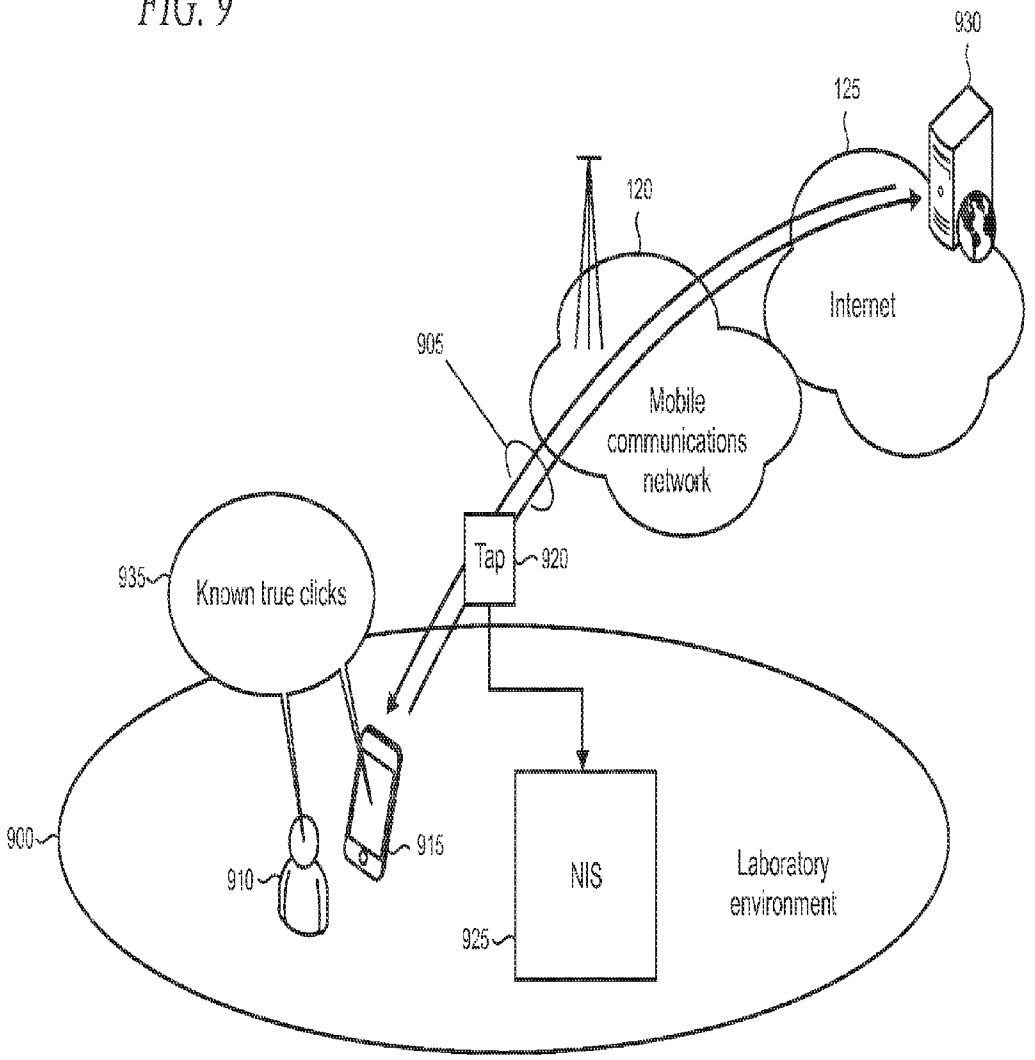

930
125
120
Internet
905
Mobile communications network
Tap
920
935
Known true clicks
900
910
915
NIS
925
Laboratory environment

FIG. 10

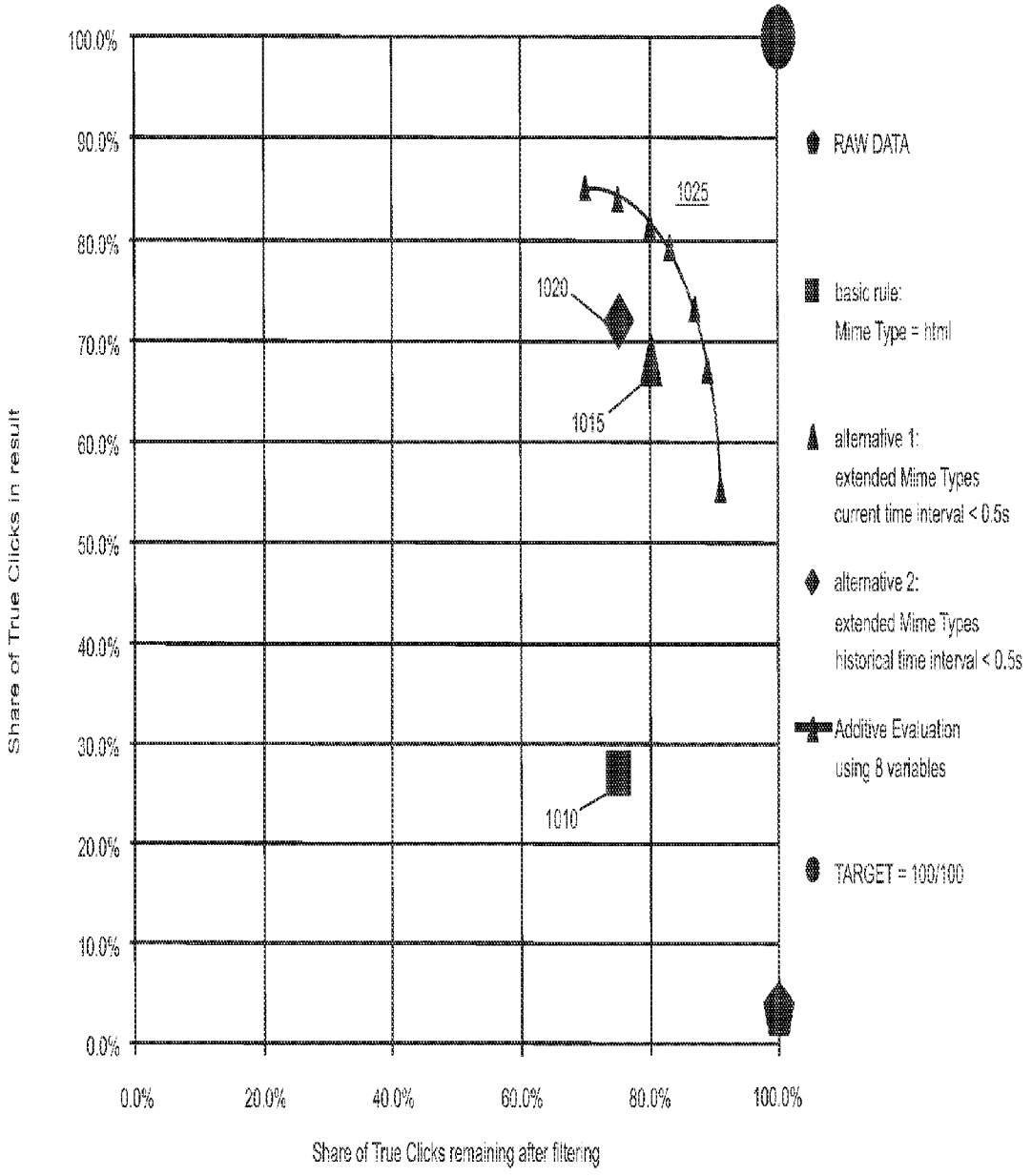

100.0%
90.0%
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
Share of True Clicks in result
0.0%
20.0%
40.0%
60.0%
80.0%
100.0%
Share of True Clicks remaining after filtering
1025
1020
1015
1010
RAW DATA
basic rule:
Mime Type = html
alternative 1:
extended Mime Types
current time interval < 0.5s
alternative 2:
extended Mime Types
historical time interval < 0.5s
Additive Evaluation
using 8 variables
TARGET = 100/100

FIG. 11

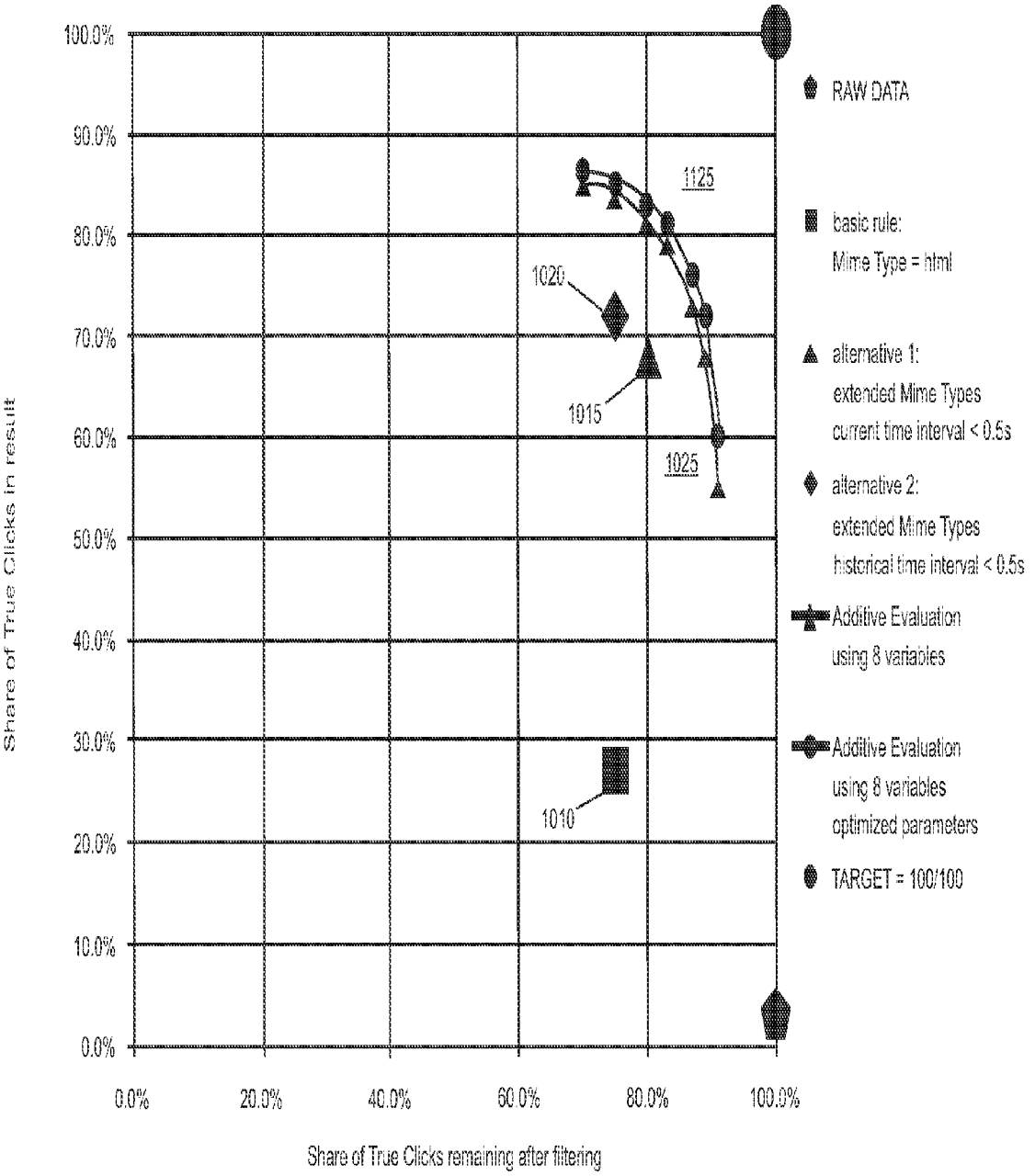
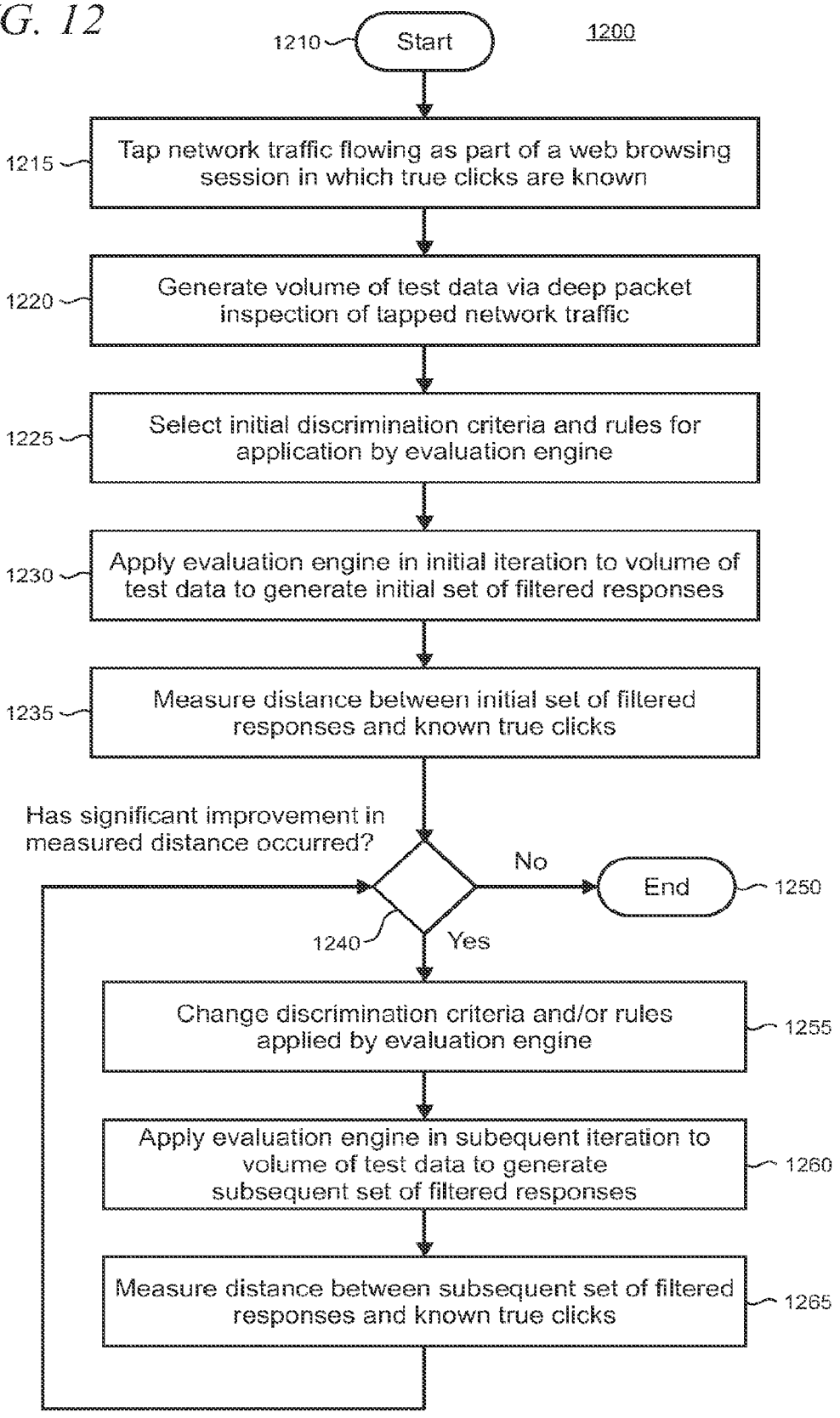

100.0%
90.0%
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
Share of True Clicks in result
0.0%
20.0%
40.0%
60.0%
80.0%
100.0%
Share of True Clicks remaining after filtering
1125
1020
1015
1025
1010
RAW DATA
basic rule:
Mime Type = html
alternative 1:
extended Mime Types
current time interval < 0.5s
alternative 2:
extended Mime Types
historical time interval < 0.5s
Additive Evaluation
using 8 variables
Additive Evaluation
using 8 variables
optimized parameters
TARGET = 100/100

FIG. 12


1200
1210
Start
1215
Tap network traffic flowing as part of a web browsing session in which true clicks are known
1220
Generate volume of test data via deep packet inspection of tapped network traffic
1225
Select initial discrimination criteria and rules for application by evaluation engine
1230
Apply evaluation engine in initial iteration to volume of test data to generate initial set of filtered responses
1235
Measure distance between initial set of filtered responses and known true clicks
Has significant improvement in measured distance occurred?
1240
No
End
1250
Yes
Change discrimination criteria and/or rules applied by evaluation engine
1255
Apply evaluation engine in subequent iteration to volume of test data to generate subsequent set of filtered responses
1260
Measure distance between subsequent set of filtered responses and known true clicks
1265

METHOD FOR GENERATING RULES AND PARAMETERS FOR ASSESSING RELEVANCE OF INFORMATION DERIVED FROM INTERNET TRAFFIC

BACKGROUND

[0001] Communication networks provide services and features to users that are increasingly important and relied upon to meet the demand for connectivity to the world at large. Communication networks, whether voice or data, are designed in view of a multitude of variables that must be carefully weighed and balanced in order to provide reliable and cost effective offerings that are often essential to maintain customer satisfaction. Accordingly, being able to analyze network activities and manage information gained from the accurate measurement of network traffic characteristics is generally important to ensure successful network operations.

[0002] This Background is provided to introduce a brief context for the Summary and Detailed Description that follow. This Background is not intended to be an aid in determining the scope of the claimed subject matter nor be viewed as limiting the claimed subject matter to implementations that solve any or all of the disadvantages or problems presented above.

SUMMARY

[0003] A method is disclosed for generating a set of optimal rules and parameters for use by an evaluation engine on a volume of information extracted from a stream of IP (Internet Protocol) packets associated with a web browsing session conducted over a network in order to filter user-initiated traffic flowing across the network from non-user-initiated traffic. Deep packet inspection ("DPI") is performed to extract the volume of information from the stream that conforms to at least one discrimination criteria. An initial iteration of application of the evaluation engine to the volume is performed by selecting initial discrimination criteria and rules for generating the filtered results and a distance between the filtered results, and known actual user-initiated traffic is measured. Subsequent iterations of application of the evaluation engine to the volume are performed by changing the discrimination criteria and/or rules until there is no significant improvement in the measured distance.

[0004] In various illustrative examples of the present method, a user engages in the web browsing session utilizing a network access device such as a mobile phone or smartphone in a laboratory environment to access web pages from servers located on the Internet over a mobile communications network using a request-response protocol such as HTTP (HyperText Transfer Protocol) or SIP (Session Initiation Protocol). In the laboratory, the user and device may be observed to ascertain the "true clicks" (i.e., responses from the server that correspond to user-initiated requests) and "false clicks" (i.e., responses that correspond to non-user-initiated requests such as those implemented through embedded scripts) that are made during the web browsing session. Discrimination criteria such as technical data, page information, or timing-based information are observed by a DPI machine to generate a volume of test data. The rules may include deterministic rules and rules implementing aggregative evaluation of the discrimination criteria (which can be weighted differently). Generation of an evaluation engine may be iterated so that discrimination criteria and rules are applied to the volume of test data in various combinations until filtered results having a maximum number of true clicks and a minimum number of false clicks are obtained.

[0005] This Summary is provided to introduce a selection of concepts in a simplified form that are further described below in the Detailed Description. This Summary is not intended to identify key features or essential features of the claimed subject matter, nor is it intended to be used as an aid in determining the scope of the claimed subject matter.

DESCRIPTION OF THE DRAWINGS

[0006] FIG. **1** shows an illustrative mobile communication network environment in which a set of optimal rules and parameters applied by an evaluation engine may be utilized;

[0007] FIG. **2** shows an illustrative web browsing session which utilizes a request-response communication protocol;

[0008] FIG. **3** shows how responses can be both user-initiated and non-user-initiated and include HTML (HyperText Markup Language) objects and embedded objects;

[0009] FIG. **4** shows an illustrative network NIS that may be located in a mobile communications network or node thereof;

[0010] FIG. **5** shows an illustrative set of variables that may be output from a deep packet inspection machine and the selection of a subset therein that are utilized as discrimination criteria in the present method;

[0011] FIG. **6** shows an illustrative taxonomy of discrimination criteria that may be utilized in the present method for generating the rules and parameters;

[0012] FIG. **7** shows an illustrative data flow from the deep packet inspection machine through an evaluation engine to produce filtered results which may be used to identify network access device user activities;

[0013] FIG. **8** shows a chart depicting an ideal target for the filtered results in which the x-axis represents the share of "true clicks" remaining after filtering and the y-axis represents the share of "true clicks" in the results;

[0014] FIG. **9** shows an illustrative laboratory environment in which known true clicks associated with a web browsing session may be obtained to generate the present rules and parameters;

[0015] FIG. **10** shows a chart depicting filtered results from the application of an evaluation engine using several different criteria;

[0016] FIG. **11** shows a chart depicting filtered results from the application of an evaluation engine that uses optimal weighting of a selected set of discrimination criteria; and

[0017] FIG. **12** shows a flowchart of an illustrative method for generating a set of optimal rules and parameters for use by an evaluation engine.

[0018] Like reference numerals indicate like elements in the drawings. Unless otherwise indicated, elements are not drawn to scale.

DETAILED DESCRIPTION

[0019] FIG. **1** shows an illustrative mobile communication network environment **100** in which a set of optimal rules and parameters applied by an evaluation engine may be utilized. It is recognized that effective analysis of network traffic can provide benefits to both network operators and users of the network (i.e., customers) by enabling, for example, the appropriate resources to be invested to ensure optimal utilization of the network's capacity and effective congestion control,

while delivering reliable and high quality service and a rich feature set to the network user. In addition, analysis of users' behaviors when accessing resources such as web pages over the network can help network providers, resource hosts, or third parties to tailor services, products, or other offerings that are responsive to the network users' wants and expectations.

[0020] As shown in FIG. **1**, a number of users $\mathbf{105}_{1, 2 \ldots N}$ of respective network access devices $\mathbf{110}_{1, 2 \ldots N}$ may access resources provided from various web servers $\mathbf{115}_{1, 2 \ldots N}$. Access is implemented, in this illustrative example, via a mobile communications network **120** that is operatively connected to the web servers **115** via the Internet **125**. It is emphasized that the present method is not necessarily limited in applicability to mobile communications network implementations and that other network types that facilitate access to the World Wide Web including local area and wide area networks, PSTNs (Public Switched Telephone Network), and the like that may incorporate both wired and wireless infrastructure may be utilized in some implementations. In this illustrative example, the mobile communications network **120** may be arranged using one of a variety of alternative networking standards such as UMTS (Universal Mobile Telecommunications System), GSM/EDGE (Global System for Mobile Communications/Enhanced Data rates for GSM Evolution), CDMA (Code Division Multiple Access), CDMA2000, or other 2G, 3G, or 4G (2^{nd}, 3^{rd}, and 4^{th} generation, respectively) wireless standards, and the like.

[0021] The network access devices **110** may include any of a variety of conventional electronic devices or information appliances that are typically portable and battery-operated and which may facilitate communications using voice and data. For example, the network access devices **110** can include mobile phones, e-mail appliances, smartphones, PDAs (personal digital assistants), ultra-mobile PCs (personal computers), tablet devices, tablet PCs, handheld game devices, digital media players, digital cameras including still and video cameras, GPSs (global positioning systems) navigation devices, pagers, or devices which combine one or more of the features of such devices. Typically, the network access devices **110** will include various capabilities such as the provisioning of a user interface that enables a user **105** to access the Internet **125** and browse and selectively interact with web pages that are served by the Web servers **115**, as representatively indicated by reference numeral **130**.

[0022] A network intelligence solution ("NIS") **135** is also provided in the environment **100** and operatively coupled to the mobile communications network **120**, or to a network node thereof (not shown) in order to access traffic that flows through the network or node and utilize an evaluation engine that may apply the optimal rules and parameters generated using the present method. In alternative implementations, the NIS **135** can be located remotely from the mobile communications network **120** and be operatively coupled to the network, or network node, using a communications link **140** over which a remote access protocol is implemented.

[0023] It is noted that performing network traffic analysis from a network-centric viewpoint can be particularly advantageous in many scenarios. For example, attempting to collect information at the client network access devices **110** can be problematic because such devices are often configured to utilize thin client applications and typically feature streamlined capabilities such as reduced processing power, memory, and storage compared to other devices that are commonly used for web browsing such as PCs. In addition, collecting data at the network advantageously enables data to be aggregated across a number of network access devices **110**, and further reduces intrusiveness and the potential for violation of personal privacy that could result from the installation of monitoring software at the client. The NIS **135** is described in more detail in the text accompanying FIG. **4** below.

[0024] FIG. **2** shows an illustrative web browsing session which utilizes a protocol such as HTTP or SIP. In this particular illustrative example, the web browsing session utilizes HTTP which is commonly referred to as a request-response protocol that is typically utilized to transfer Web files. Each transfer consists of file requests $\mathbf{205}_{1, 2 \ldots N}$ for pages or objects from a browser application executing on the network access device **110** to a server **115** and corresponding responses $\mathbf{210}_{1, 2 \ldots N}$ from the server. Thus, at a high level, the user **105** interacts with a browser to request, for example, a URL (Uniform Resource Locator) to identify a site of interest, then the browser requests the page from the server **115**. When receiving the page, the browser parses it to find all of the component objects such as images, sounds, scripts, etc., and then makes requests to download these objects from the server **115**.

[0025] As shown in FIG. **3**, a webpage is primarily an HTML (HyperText Markup Language) object (representatively indicated by reference numeral **305**) typically having a content type of text/html with links to other objects $\mathbf{310}_{1 \ldots N}$ in it as embedded objects (images, sounds, scripts, etc.). A webpage may accordingly be generated either in response to a direct user-initiated request (also termed a "true click"), as indicated by reference numeral **315**, or due to a non-user-initiated request (also termed a "false click"), as indicated by reference numeral **320** via execution, for example, of an embedded script at the client network access device **110**. Such script execution can result in a substantial amount of network traffic to be automatically generated and to flow to the network access device **110** through the mobile communications network **120**. For example, a visit at the news site CNN.com with 5 page views will create 650 HTTP events in which 100 of them are HTML.

[0026] FIG. **4** shows details of the NIS **135** which is arranged, in this illustrative example, to identify user-initiated traffic and distinguish it from non-user-initiated traffic by examining network traffic through the mobile communications network **120**. The NIS **135** is typically configured as one or more software applications or code sets that are operative on a computing platform such as a server **405** or distributed computing system. In alternative implementations, the NIS **135** can be arranged using hardware and/or firmware, or various combinations of hardware, firmware, or software as may be needed to meet the requirements of a particular usage scenario.

[0027] The NIS **135** comprises a deep packet inspection ("DPI") machine **410** and an evaluation engine **415** that writes to a database **420**. The database **420** may be accessed, manipulated, and queried to perform analysis of the usage of the mobile communications network **120**, as indicated by reference numeral **425** in FIG. **4**. DPI machines are known, and commercially available examples include the ixMachine produced by Qosmos SA.

[0028] As shown, traffic typically in the form of IP packets **430** flowing through the mobile communications network **120**, or a node of the network, are captured via a tap **435** in a packet capture component **440** of the DPI machine **410**. An engine **445** takes the captured IP packets to extract various

types of information, as indicated by reference numeral **450**, and filter and/or classify the IP traffic **430**, as indicated by reference numeral **455**. An information delivery component **460** of the DPI machine **410** then outputs the data generated by the DPI engine **445** to the evaluation engine **415**, as shown. The evaluation engine **415** uses various evaluation rules **465** through the application of one or more of the discrimination criteria **470** in various combinations in order to identify user-initiated traffic in the IP traffic **430**.

[0029] FIG. **5** shows an illustrative set of variables **505** that may be output from the DPI machine **410** (FIG. **4**) and the selection of a subset therein that are utilized as discrimination criteria **470** in the present method. As shown, the DPI machine **410** has the capability to produce a very large set of variables that can be captured from the IP traffic **430** (FIG. **4**). These variables illustratively include traffic attributes **510**, application content **515**, content attributes **520**, session detail records ("SDRs") **525**, and metadata attributes **530** among other variables. In accordance with the principles of the present method for generating optimal rules and parameters, it is noted that a particular subset of the myriad of available variables **505** is particularly well-suited for use as discrimination criteria **470**. This includes technical data **540**, page information **545**, and timing-based information **550** which arc then applied using the rules **465** by the evaluation engine **415** to identify user-initiated request/response pairs **555**.

[0030] The selection of the technical data **540**, page information **545**, and timing-based information **550** may be implemented, for example, by executing the appropriate code in the DPI machine Turning again to FIG. **4**, for example, software code may execute in a configuration and control layer **475** in the DPI machine **410** to select the discrimination criteria from among the variables that are available for output by the engine **445** in the DPI machine **410**.

[0031] FIG. **6** shows an illustrative taxonomy **600** of discrimination criteria **470** that may be applied by the rules **465** (FIG. **4**) in the evaluation engine **415**. It is emphasized that the taxonomy **600** is intended to be illustrative of the variables that have been determined to be good candidates to identify user-initiated request/response pairs in many typical applications. However, the variables illustrated in taxonomy **600** should not be viewed as an exhaustive listing of all suitable variables. As shown, the technical data **540** illustratively includes MIME (Multipurpose Internet Mail Extension) type **605** such as text/html, image/jpeg, application/x-javascript, xhtml+xml, and the like. The technical data **540** further includes response codes **610** (i.e., status codes) from a Web server **115** where, for example, response codes 200-299 indicate OK, codes 301-304 indicate redirection, and codes 400-999 indicate errors.

[0032] The page information **545** illustratively includes file extensions **615** such as .jpg, .bmp, .gif, .htm, .js, etc. Referrer information **620** may include web pages without a referrer (i.e., where a referrer identifies, from the point of view of a webpage, the address or URL of the resource which links to it). The page information **545** may further include page titles and meta-tags **625** where the meta-tags may include, for example, search words, and also include a URI (Uniform Resource Identifier) to a home page **630**. Page information **545** may further include an historical average number of requests **635** that are received at a particular server **115**. Variables included in the page information **545** also include pages both with and without a response having cookies (including third-party cookies), as indicated by reference numeral **640**, and pages both with and without a request for a favorite icon (also termed a "favicon"), as indicated by reference numeral **645**.

[0033] The timing-based information **550** illustratively includes the time interval between a current request (e.g., request **205** in FIG. **2**) to a former (i.e., preceding) request, as indicated by reference numeral **650**. The timing-based information **550** may also include the time interval between a current request and a referrer, as indicated by reference numeral **655**.

[0034] Under the HTTP 1.1 standard, multiple successive requests may be written out to a single network socket without waiting for a corresponding response from the remote server in a process known as "pipelining." The requestor (e.g., the browser) then waits for the responses to arrive in the order in which they were requested. The pipelining of requests can result in a significant improvement in page loading times, especially over high latency connections. The time interval between a current request and a request in the same base flow when using the pipelining technique, as indicated by reference numeral **670** may also be included in the timing-based information **550**. The timing-based information **550** may further include observations of the history of the time intervals between requests **675**, as well as the historical time interval to a referrer **680**.

[0035] As noted above, the evaluation rules **465** used by the engine **415** (FIG. **4**) are applied to the network traffic using the discrimination criteria **470** in order to identify user-initiated requests and corresponding server responses and further distinguish those requests/responses from non-user-initiated responses that may be generated, for example, through execution of embedded scripts. In other words, as shown in FIG. **7**, data **705** generated from the DPI engine **445** (FIG. **4**) is filtered through the application of the evaluation rules to the discrimination criteria, as indicated at reference numeral **710**, to produce a set of filtered results **715**. The optimal target of such filtering would be a one-to-one correspondence between the filtered results **715** and the user-initiated responses.

[0036] FIG. **8** depicts a chart **800** which expresses this target graphically in which the x-axis indicates the share of true clicks remaining after filtering and the y-axis represents the share of true clicks in the filtered results. The target **805** is at 100% on the x-axis and 100% on the y-axis which means that no true clicks are missed (i.e. the filtered results **715** are not under-inclusive of true clicks) and only true clicks are included (i.e., the filtered results are not over-inclusive to include false clicks).

[0037] The evaluation engine **415** can be applied to a volume of test data that may be obtained under controlled conditions, for example, in a laboratory environment **900** as shown in FIG. **9**. In the laboratory environment **900**, IP packets **905** associated with a web browsing session of a known individual user **910** and network access device **915** may be tapped via a tap **920** by an NIS **925** that is co-located in the laboratory or otherwise provided with remote access to it. The NIS **925** can be arranged in a similar manner to the NIS **135** shown in FIGS. **1** and **4** and described in the accompanying text.

[0038] As shown in FIG. **9**, the network access device **915** may access one or more known servers, as representatively indicated by reference numeral **930**, via the mobile communications network **120** and the Internet **125**. However, in some implementations, characteristics of the mobile communications network **120** and the Internet **125** can be simulated in the

laboratory environment. Alternatively, other types of networks such as local area networks or virtual private networks may be desirably utilized.

[0039] During the web browsing session in the laboratory environment **900**, observations of the user **910** and/or the network access device **915** may be made in order to obtain a set of known true clicks **935** that may be used to define parameters associated with the ideal target **805** shown in FIG. **8**. In some cases, information describing the behaviors and actions of the known server **930**, such as code that executes on the server **930**, may be optionally utilized to further enhance definition of the target **805**, or for other purposes in the laboratory.

[0040] In one illustrative example of data collection in the laboratory environment **900**, several volumes of test data were generated for web browsing sessions on several websites with many page views. Utilization of a DPI machine created a large amount of request and response objects showing the timing of the request and response, the URL/URI of the referrer, the MIME type, and the response code. Additional information was added to the objects for test purposes including the true click information, page titles, and meta tags.

[0041] In the laboratory environment **900** various evaluation rules and discrimination criteria may be tested alone or in different combinations to generate filtered results from the volume of test data that can be compared against the set of known true clicks **935** to assess whether a given evaluation engine applying such rules and criteria provides results that are acceptably close to the target **805**. The evaluation rules may encompass a range of rules and include relatively straightforward deterministic rules as well as more complex rules that utilize, for example, the aggregation of evaluations of a plurality of discrimination criteria (i.e., variables), where the evaluations can be weighted differently. The aggregation may be performed, for example, on an additive or multiplicative basis.

[0042] An example of application of an illustrative basic deterministic rule set is one that includes a response in the filtered results if the MIME type=text/html, and the response code=2xx (i.e., indicating that the corresponding request was successfully received, understood, and accepted), while excluding responses with file extensions like .jpg, .bmp, .gif, .js, and the like. When applied in an evaluation engine on the volume of test data having known true clicks, the performance of the engine is fairly poor with 25% of true clicks missed from the filtered results and many false clicks included, yielding a result of 75% on the x-axis and 28% on the y-axis, as indicated by the symbol **1010** in FIG. **10**.

[0043] An illustrative first alternative deterministic rule set using current time intervals can be utilized by an evaluation engine as follows: the application of rule 1 results in the inclusion of an object in a response in the filtered results if the object is determined to belong to a group MIME type=text/html (or a comparable group such as xhtlm, xml, plain/text, etc.). Rule 2 includes a response object in the results when a server response code=2xx. Rule 3 excludes an object having a particular file extension such as .jpg, .bmp, .gif, .js, and the like. Rule 4 excludes an object if the time interval to a former request is less than 0.5 seconds. Application of this first alternative rule set to the volume of test data yields 20% of true clicks missed from the filtered results and comparably fewer false clicks included for a result of 80% on the x-axis and 68% on the y-axis as indicated by the symbol **1015** in FIG. **10**.

[0044] An illustrative second alternative deterministic rule set using historical time intervals can also be utilized by an evaluation engine as follows: Rules 1-3 are the same as in above example. Rule 4 excludes an object from the results if the historical time interval to a former request was, in 70% of the cases, less than 0.5 seconds. Application of this second alternative rule set to the volume of test data yields 25% of true clicks missed from the filtered results and comparably fewer false clicks included for a result of 75% on the x-axis and 72% on the y-axis as indicated by the symbol **1020** in FIG. **10**.

[0045] An example of a more complex rule set illustratively includes an evaluation of an object based on the aggregative evaluation of several discrimination criteria. This rule set relies upon the observation that some MIME types and file extensions are more likely to be associated with user-initiated actions, others are less likely, and some are definitely not associated. In addition, objects without a referrer and objects that are referrers for other objects are more likely to be associated with user-initiated actions. And, objects that appear with a high time interval or show a historically high median time interval are more likely to be associated with user-initiated actions. Here, each subjective weighting is applied (and expressed as points) to the set of eight discrimination criteria below:

[0046] +10 if MIME type=text/html; +5 if MIME type=xml; −50 if MIME type=jpg, gig, bmp, etc.

[0047] +5 if home page (i.e., HTTP URL path=/)

[0048] +5 if a current time interval to former request is above 0.5 sec or +10 if above 2 sec.

[0049] +5 if an historical time interval to a former request is on average above 0.5 sec or +10 if above 2 sec.

[0050] −10 if the current time interval in the same base flow is below 0.1 sec.

[0051] +3 if an object has no referrer and/or is the object is a referrer of other events.

[0052] +3 if the object has a title or meta tags

[0053] +1 if the object requests cookies and/or favorite icons

[0054] This rule set enables calculation of the consequences of specific threshold values. It is observed that an increase of the threshold value will increase the exclusion rate of false clicks, but also increases the probability of excluding true clicks. Application of this complex rule set to the volume of test data yields results that vary between 91/55 (percentages on the respective x-axis and y-axis) and 70/85 depending on the particular threshold values selected, as shown by symbols **1025** in FIG. **10**. The aggregation of values may be changed in several ways, for example, by adding or replacing criteria, changing the weights of a single criterion, or by changing the method of aggregation. Accordingly, the complex rule set can be further refined using the additive aggregation expression:

$$_{16B}p = \sum_{i=1}^{n} b_i * v_i$$

where p is the probability that an object is associated with a true click, v is the variable discrimination criterion, and b is the weight. Multiplicative aggregation may be alternatively implemented in some cases according to:

$$18B\,p = \prod_{i=1}^{n} v_i^{b^i}$$

[0055] When using the additive aggregation expression, optimized weights for the eight discrimination criteria listed above can be calculated for a basic data set using standard dummy regression. Application of an evaluation engine using the optimized weights demonstrates improved filtering performance for various threshold values as shown by the symbols **1125** in FIG. **11**.

[0056] FIG. **12** shows a flowchart of an illustrative method **1200** for generating a set of optimal rules and parameters for use by an evaluation engine. The method begins at block **1210**. At block **1215**, traffic flowing as part of a web browsing session is tapped to collect IP packets. As noted above, such collection can typically take place in a laboratory environment where conditions are controlled and the user is known so that true clicks can be directly observed. This laboratory environment can be expected to differ from the field environment in which the evaluation engine utilized in the method **1200** is deployed where the users **105** can typically be expected to be members of the public who are customers of the mobile communications network.

[0057] In such a field environment, the data collected and utilized by the NIS **135** (FIGS. **1** and **4**), or portions thereof can be anonymized to remove identifying information from the data, for example, to ensure that privacy of the network access device users is maintained. Other techniques may also be optionally utilized in some implementations to further enhance privacy including, for example, providing notification to the users **105** that certain anonymized data may be collected and utilized to enhance network performance or improve the variety of features and services that may be offered to users in the future, and providing an opportunity to opt out (or opt in) to participation in the collection.

[0058] Anonymization may be implemented by encrypting portions or all of the tapped network traffic to obscure information from which the network access device users' identities or data that could be used to obtain their identities might otherwise be determined. In some cases, the encrypted data may include a unique "anonymizing" identifier that can be correlated to unencrypted traffic data extracted from those packets associated with a corresponding user **105**. This anonymizing process allows mobile communications network use of any individual user to be differentiated from the network use of all other users on a completely anonymous basis—that is, without referencing any personal identity information (e.g., name, address, telephone number, account number, etc.) of the user.

[0059] A volume of test data is generated via deep packet inspection of the tapped network traffic at block **1220** in FIG. **12**. An initial iteration of application of an evaluation engine is performed, at block **1225**, by selecting initial discrimination criteria and rules. At block **1230**, the initial iteration of application of the evaluation engine to a volume of test data is performed to generate an initial set of filtered responses. At block **1235**, a measurement of the distance between the initial set of filtered responses and known actual user-initiated traffic (i.e., known true clicks) is performed. Decision block **1240** is skipped after the initial set of filtered responses is generated and control passes to block **1255**.

[0060] At block **1255**, the discrimination criteria and/or rules applied by the evaluation engine are changed. The evaluation engine is applied, in a subsequent iteration, to a volume of test data to generate a subsequent set of filtered responses at block **1260**. At block **1265**, a measurement is made of the distance between the subsequent set of filtered responses and the known user-initiated traffic (i.e., known true clicks). Control is passed to decision block **1240** where a determination is made to continue to iterate the method steps **1255**, **1260**, and **1265** or end the method **1200**. If no significant improvement in the measured distance has occurred, then the method **1200** ends at block **1250**.

[0061] Although the subject matter has been described in language specific to structural features and/or methodological acts, it is to be understood that the subject matter defined in the appended claims is not necessarily limited to the specific features or acts described above. Rather, the specific features and acts described above are disclosed as example forms of implementing the claims.

1. (canceled)
2. (canceled)
3. (canceled)
4. (canceled)
5. (canceled)
6. (canceled)
7. (canceled)
8. (canceled)
9. (canceled)
10. (canceled)
11. (canceled)
12. (canceled)
13. (canceled)
14. (canceled)
15. (canceled)
16. (canceled)
17. (canceled)
18. (canceled)
19. (canceled)
20. (canceled)

21. A computer-implemented method for distinguishing between true clicks and false clicks in respective web browsing sessions between corresponding network access devices and a remote server, the method comprising the steps of:

applying deep packet inspection to a stream of IP packets utilized in the web-browsing sessions to extract selected information from the IP packets according to discrimination criteria, the discrimination criteria including at least one of technical data, page information, or timing-based information;

operating an evaluation engine incorporating at least one rule or criterion modifiable to progress toward a one to one correspondence between a number of modeled true clicks and a number of true clicks known from selected information extracted from a sample stream of IP packets according to the discrimination criteria; and

applying rules obtained by operation of the evaluation engine to information extracted from web browsing sessions between the network access devices and a remote server.

22. The method of claim **21**, wherein web-browsing sessions originating between the network access devices and a remote server include hypertext transfer protocol (http) information requests.

23. The method of claim **21**, wherein web-browsing sessions originating between the network access devices and a remote server include information requests and wherein the at least one rule or criterion applied during operation of the evaluation engine includes a requirement that the file type specified by a response to an information request derived from the sample stream have a text/html, xhtml, xml, or plain/text extension.

* * * * *