



- (51) **International Patent Classification:**
G06F 15/16 (2006.01)
- (21) **International Application Number:**
PCT/US2013/039738
- (22) **International Filing Date:**
6 May 2013 (06.05.2013)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (30) **Priority Data:**
13/791,274 8 March 2013 (08.03.2013) US
- (71) **Applicant (for all designated States except US):**
GOOGLE INC. [US/US]; 1600 Amphitheatre Parkway,
Mountain View, California 94043 (US).
- (72) **Inventor; and**
- (71) **Applicant (for US only):** **YIM, Keun Soo** [KR/US]; 1180
Lochinvar Avenue, #124, Sunnyvale, California 94087
(US).
- (74) **Agents:** **DRAGSETH, John A.** et al.; Fish & Richardson
P.C., P.O. Box 1022, Minneapolis, Minnesota 55440-1022
(US).
- (81) **Designated States (unless otherwise indicated, for every
kind of national protection available):** AE, AG, AL, AM,
AO, AT, AU, AZ, BA, BB, BG, BH, BN, 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, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU,
RW, 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) **Designated States (unless otherwise indicated, for every
kind of regional protection available):** ARIPO (BW, GH,
GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ,
UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ,
TM), European (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).

Declarations under Rule 4.17:

- as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))
- as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))

Published:

- with international search report (Art. 21(3))

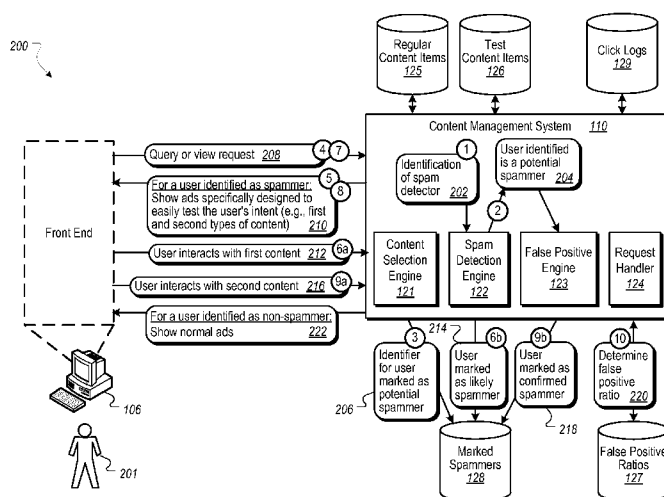
(54) **Title:** DETERMINING A FALSE POSITIVE RATIO OF A SPAM DETECTOR

FIG. 2

(57) **Abstract:** Methods, systems, and apparatus include computer programs encoded on a computer-readable storage medium, including a method for providing content. The method includes marking a particular user is a spammer, receiving a request for content from the user, providing a first type of content item that tests whether the user is a likely spammer, and when the user interacts with the first type of content item, marking the user as a likely spammer. The method further includes receiving a subsequent request for content, providing the user with a second different type of content item, and when the user interacts with the second different type of content item, marking the user as a spammer. The method further includes determining a false positive ratio of the spam detector based on a number of times users are marked as potential spammers versus the number of times that users are ultimately marked as spammers.

DETERMINING A FALSE POSITIVE RATIO OF A SPAM DETECTOR

CROSS-REFERENCE TO RELATED APPLICATION

[001] This application claims priority to U.S. Application Serial No. 13/791,274, filed on March 8, 2013, entitled DETERMINING A FALSE POSITIVE RATIO OF A SPAM DETECTOR, the disclosure of which is incorporated herein by reference.

BACKGROUND

[002] This specification relates to information presentation.

[003] The Internet provides access to a wide variety of resources. For example, video and/or audio files, as well as webpages for particular subjects or particular news articles, are accessible over the Internet. Access to these resources presents opportunities for other content (e.g., advertisements) to be provided with the resources. For example, a webpage can include slots in which content can be presented. These slots can be defined in the webpage or defined for presentation with a webpage, for example, along with search results.

[004] Content slots can be allocated to content sponsors as part of a reservation system, or in an auction. For example, content sponsors can provide bids specifying amounts that the sponsors are respectively willing to pay for presentation of their content. In turn, an auction can be run, and the slots can be allocated to sponsors according, among other things, to their bids and/or the relevance of the sponsored content to content presented on a page hosting the slot or a request that is received for the sponsored content. The content can then be provided to the user on any devices associated with the user such as a personal computer (PC), a smartphone, a laptop computer, a tablet computer, or some other user device.

[005] Content sponsors can be charged when their content is presented to users. In some cases, content delivery systems can use spam detectors to determine when content is being requested by (or provided to) spammers as opposed to actual users.

SUMMARY

[006] In general, one innovative aspect of the subject matter described in this specification can be implemented in methods that include a computer-implemented method for identifying a spam detector, the spam detector operable to determine when an interaction with a content item by a user is true or is more likely to be spam. The method

further includes receiving an input that a particular user is a spammer. The method further includes recording an identifier for the user and marking the user as a potential spammer. The method further includes receiving a request for content from a potential spammer. The method further includes providing the potential spammer with a first type of content item that tests whether the potential spammer is a likely spammer. The method further includes, when the potential spammer interacts with the first type of content item, marking the potential spammer as a likely spammer. The method further includes receiving a subsequent request for content from the likely spammer. The method further includes providing the likely spammer with a second different type of content item that tests whether the likely spammer is a spammer, where selection of the second different type of content item, once presented, represents a high probability that the likely spammer is a spammer. The method further includes, when the likely spammer interacts with the second different type of content item, marking the likely spammer as a spammer. The method further includes determining a false positive ratio of the spam detector based at least in part on a number of times users are marked as potential spammers versus the number of times that users are ultimately marked as spammers.

[007] These and other implementations can each optionally include one or more of the following features. The first type of content item can be an interstitial advertisement. The tests associated with the first and second different types of content items can use different probabilities. The first type of content item can be a blinking advertisement including at least first and second states, and the first state can be a state in which the blinking advertisement is visible to the user, and the second state can be a state in which the blinking advertisement is not visible to the user. The first type of content item can be a hidden advertisement that does not degrade a user experience. The first type of content item can be a content item that includes a changeable selection area, and the method can further include determining an interaction latency when interacting with the changeable selection area. The first type of content item can be a content item that includes a fixed user-visible selection area, and the method can further include identifying an interaction location for where an entity interacted with the fixed selection area so as to be able to infer whether the interaction was associated with a human user or a machine, and the method can further include determining interaction locations that are likely associated with human interaction and interaction locations that are unlikely associated with human interactions. The second different type of content item can be selected from the group comprising a junk advertisement, an empty advertisement, a hidden advertisement, an

advertisement that is in a different language than that associated with the user, or a mismatched advertisement. Receiving an input that a particular user is a spammer can include receiving input from the spam detector that the spam detector has determined that the user is a spammer. Determining a false positive ratio can further include determining a ratio of users determined to be non-spammers due to not being marked as spammers and a sum of the users determined to be non-spammers and the users determined to be spammers. The method can further include dividing users into multiple groups, providing first and/or second types of content items to users with different probabilities based on their respective groups, and using statistical analysis of false positives of the multiple groups to determine spamminess of entire groups. The method can further include adjusting the spam detector based at least in part on the determined false positive ratio. The method can further include repeating the method after a predetermined time or when a predetermined condition is met in order to determine a new false positive ratio. The method can further include adjusting a click-through amount for an advertiser for a given campaign based at least in part on the false positive ratio. The first type of content can include a low click-through ratio advertisement, and upon detecting an interaction with the low click-through ratio advertisement, labeling the user a potential spammer.

[008] In general, another innovative aspect of the subject matter described in this specification can be implemented in methods that include another computer-implemented method for providing creatives. The method includes receiving an indication that an interaction associated with a user is determined to be spam, where spam represents a false interaction with a content item. The method further includes saving an identifier associated with the user and marking the user as a potential spammer. The method further includes receiving a request for content from the potential spammer. The method further includes providing the potential spammer with a first type of content item in response to the request or in lieu of directing a user to a landing page associated with the advertisement that tests whether the potential spammer is a likely spammer. The method further includes, when the potential spammer interacts with the first type of content item, marking the potential spammer as a likely spammer. The method further includes, thereafter, receiving a subsequent request for content from the likely spammer. The method further includes providing the likely spammer with a second different type of content item in response to the request or in lieu of directing a user to a landing page associated with the second advertisement, the second different type of content item being one that tests whether the likely spammer is a spammer, where selection of the second

different type of content item, once presented, represents a high probability that the likely spammer is a spammer. The method further includes, when the likely spammer interacts with the second different type of content item, marking the likely spammer as a spammer. The method further includes storing a final determination for the user based at least in part on the interaction with the second different type of content item by the user.

[009] In general, another innovative aspect of the subject matter described in this specification can be implemented in computer program products that include a computer program product tangibly embodied in a computer-readable storage device and comprising instructions. The instructions, when executed by one or more processors, cause the processor to: identify a spam detector, the spam detector operable to determine when an interaction with a content item by a user is true or is more likely to be spam; receive an input that a particular user is a spammer; record an identifier for the user and marking the user as a potential spammer; receive a request for content from a potential spammer; provide the potential spammer with a first type of content item that tests whether the potential spammer is a likely spammer; when the potential spammer interacts with the first type of content item, mark the potential spammer as a likely spammer; receive a subsequent request for content from the likely spammer; provide the likely spammer with a second different type of content item that tests whether the likely spammer is a spammer, where selection of the second different type of content item, once presented, represents a high probability that the likely spammer is a spammer; when the likely spammer interacts with the second different type of content item, mark the likely spammer as a spammer; and determine a false positive ratio of the spam detector based at least in part on a number of times users are marked as potential spammers versus the number of times that users are ultimately marked as spammers.

[0010] These and other implementations can each optionally include one or more of the following features. The first type of content item can be a hidden advertisement that does not degrade a user experience.

[0011] In general, another innovative aspect of the subject matter described in this specification can be implemented in systems, including a system comprising one or more processors and one or more memory elements including instructions. The instructions, when executed, cause the one or more processors to: identify a spam detector, the spam detector operable to determine when an interaction with a content item by a user is true or is more likely to be spam; receive an input that a particular user is a spammer; record an identifier for the user and marking the user as a potential spammer; receive a request for

content from a potential spammer; provide the potential spammer with a first type of content item that tests whether the potential spammer is a likely spammer; when the potential spammer interacts with the first type of content item, mark the potential spammer as a likely spammer; receive a subsequent request for content from the likely spammer; provide the likely spammer with a second different type of content item that tests whether the likely spammer is a spammer, where selection of the second different type of content item, once presented, represents a high probability that the likely spammer is a spammer; when the likely spammer interacts with the second different type of content item, mark the likely spammer as a spammer; and determine a false positive ratio of the spam detector based at least in part on a number of times users are marked as potential spammers versus the number of times that users are ultimately marked as spammers.

[0012] These and other implementations can each optionally include one or more of the following features. The type of content item can be a hidden advertisement that does not degrade a user experience.

[0013] Particular implementations may realize none, one or more of the following advantages. For example, the techniques described herein can be used to validate a spam detection ratio of a production spam detector, including determining a more accurate false positive ratio. The portion of uncharged clicks can be reduced that are mistakenly excluded from being charged to advertisers due to false positives in a spam detector. A false positive ratio for a spam detector can be decreased significantly. Spammers may be more accurately able to be detected. Spammers may be demotivated by providing different content (e.g., junk advertisements) once detected. Better spam detection can prevent advertisers from having to run their own live experiments, e.g., including the use of junk and/or interstitial pages, to estimate the spam ratio of advertisement networks.

[0014] The details of one or more implementations of the subject matter described in this specification are set forth in the accompanying drawings and the description below. Other features, aspects, and advantages of the subject matter will become apparent from the description, the drawings, and the claims.

BRIEF DESCRIPTION OF THE DRAWINGS

[0015] FIG. 1 is a block diagram of an example environment for delivering content.

[0016] FIG. 2 shows an example system for determining a false positive ratio in a spam detector.

[0017] FIG 3 is a block diagram showing example characteristics of different levels of tests performed over time to detect spammers.

[0018] FIG. 4 is a flowchart of an example process for determining a false positive ratio of a spam detector.

[0019] FIG. 5 is a flowchart of an example process for classifying users as spammers.

[0020] FIG. 6 is a block diagram of an example computer system that can be used to implement the methods, systems and processes described in this disclosure.

[0021] Like reference numbers and designations in the various drawings indicate like elements.

DETAILED DESCRIPTION

[0022] This document describes methods, processes and systems for determining false positive ratios in a spam detector associated, e.g., with user interactions with online content. For example, content sponsors can sponsor content such as advertisements that are provided to users. The users may be actual users, or they may be spammers, such as automated systems including robots (or bots), malicious interlopers or automated computers, or systems that pose as actual users. Under some arrangements, content sponsors are not charged for presentations of their content to spammers. Techniques for determining spammers can sometimes provide false positives, e.g., erroneously identifying a user as a spammer when the user is actually a true user. This can cause content delivery networks/servers to lose revenue that would otherwise be charged to content sponsors due to false positives. The techniques herein can identify false positive ratios that can be used, for example, to more accurately charge content providers for content provided to actual users. The false positive ratios can be identified, for example, among a group of suspected spammers, e.g., a group of users identified as spammers but for which a false positive ratio exists.

[0023] In some implementations, advertisers may run experiments on specific advertisement networks to estimate spam click ratios. For example, some techniques may provide an interstitial page when an advertisement is clicked. In some implementations, interstitial pages (or advertisements) can include hidden pages (or advertisements) that do not degrade user experience. A Bayesian model can be used to estimate the spam click ratio by using the measured difference between original advertisements and advertisements with an interstitial page. However, running such experiments can largely

degrade the quality of user experience, e.g., for the target users that are randomly chosen in a certain population (e.g., 0.01% of all users).

[0024] In some implementations, the overall user experience can be improved by taking a different approach to spam detection. For example, when a spam detector identifies a user as a suspected spammer (e.g., identified by IP address, cookie, header order, user agent or some other identifier), a false positive detector can mark the spammer as a potential spammer and monitor that particular user. Then, if the potential spammer subsequently generates a request for content or interacts with content (e.g., clicks on an ad), then a content item can be displayed that tests whether the user is a spammer or a non-spammer. This can be a first-level test, e.g., using a first type of content. In some implementations, the first type of content item can be a hidden advertisement that does not degrade user experience.

[0025] In some implementations, in a first level of testing, for example, strategies can include not adversely affecting the user experience but still effectively identifying non-spammers. For example, blinking advertisements can be presented, and the click time can be evaluated, e.g., to determine a correlation between a click event time and the blinking period. If an apparent user clicks at the time during which the advertisement disappears or is not displayed, then the user can be marked as a likely spammer. If the user passes the tests that are provided (e.g., for a sufficient number of times), then the user can be marked as a non-spammer.

[0026] At a second level of testing, when a previously-identified likely spammer subsequently sends a request for content or otherwise interacts with content (e.g., submits a query, interacts with a content item), a different kind of content item can be displayed that tests whether the user is a spammer. In this second level of testing, strategies can include maximizing the detection ratio of spammers, even at the cost of adversely affecting the user experience. For example, likely spammers can be presented with “junk” advertisements such as empty advertisements, advertisements using a different language than the user’s own language, or mismatched advertisements (e.g., the advertisement’s image does not match the text). If the user interacts with (e.g., clicks on) a junk advertisement, for example, then the user can be marked as a confirmed spammer. Other types of content can be shown to likely spammers, and other types or categories of spammers can exist.

[0027] In some implementations, users who are identified as spammers can be provided with different content than the content that is provided to non-spammers. For

example, in click-to-call functions, users identified as spammers can be directed to a different phone number or put on hold. In another example, if a user identified as a spammer clicks on an advertisement, the user can be directed to a different landing page or presented with interstitial pages or advertisements that can help to verify (or contradict) that the user is a spammer. In some implementations, when click-to-call advertisements are provided, ways that can be used to detect a spammer include analyzing the length or duration of the call, e.g., as compared to typical lengths or durations of calls by non-spammers. The detection can determine, for example, if the user is a confirmed spammer, or is found to be a non-spammer.

[0028] In some implementations, information about which users are spammers/non-spammers can expire over time. For example, time limits can be set under which users designated as spammers are blacklisted. Further, testing can show, for example, that an IP address generating spammy click patterns yesterday (and therefore designated as a spammer) is no longer generating the same spammy click patterns today (and may therefore be a non-spammer).

[0029] Over time, information about the number of users who are spammers versus non-spammers can be used for various purposes. In some implementations, clicks generated by spammers and non-spammers can be counted. For example, when the numbers of potential and likely spammers are more accurately known, more accurate false positive ratios can be calculated based on numbers of spammer clicks and non-spammer clicks, or in other ways.

[0030] In some implementations, once the false positive ratio of a spam detector is known, the use of techniques for determining false positive ratio can be suspended, e.g., in order to reduce overhead and to return user experience to pre-test levels. Over time, the techniques can be repeated periodically, such as for a time period that is determined experimentally to adequately maintain false positive ratios that are sufficiently accurate.

[0031] FIG. 1 is a block diagram of an example environment 100 for delivering content. The example environment 100 includes a content management system 110 for selecting and providing content in response to requests for content. The example environment 100 includes a network 102, such as a local area network (LAN), a wide area network (WAN), the Internet, or a combination thereof. The network 102 connects websites 104, user devices 106, content sponsors 108 (e.g., advertisers), publishers 109, and the content management system 110 (e.g., including at least one spam detector). The

example environment 100 may include many thousands of websites 104, user devices 106, content sponsors 108 and publishers 109.

[0032] A website 104 includes one or more resources 105 associated with a domain name and hosted by one or more servers. An example website is a collection of webpages formatted in HTML that can contain text, images, multimedia content, and programming elements, such as scripts. Each website 104 can be maintained by a content publisher, which is an entity that controls, manages and/or owns the website 104.

[0033] A resource 105 can be any data that can be provided over the network 102. A resource 105 can be identified by a resource address that is associated with the resource 105. Resources include HTML pages, word processing documents, portable document format (PDF) documents, images, video, and news feed sources, to name only a few. The resources can include content, such as words, phrases, images, video and sounds, that may include embedded information (such as meta-information hyperlinks) and/or embedded instructions (such as JavaScript™ scripts).

[0034] A user device 106 is an electronic device that is under control of a user and is capable of requesting and receiving resources over the network 102. Example user devices 106 include personal computers (PCs), televisions with one or more processors embedded therein or coupled thereto, set-top boxes, mobile communication devices (e.g., smartphones), tablet computers and other devices that can send and receive data over the network 102. A user device 106 typically includes one or more user applications, such as a web browser, to facilitate the sending and receiving of data over the network 102.

[0035] A user device 106 can request resources 105 from a website 104. In turn, data representing the resource 105 can be provided to the user device 106 for presentation by the user device 106. The data representing the resource 105 can also include data specifying a portion of the resource or a portion of a user display, such as a presentation location of a pop-up window or a slot of a third-party content site or webpage, in which content can be presented. These specified portions of the resource or user display are referred to as slots (e.g., ad slots).

[0036] To facilitate searching of these resources, the environment 100 can include a search system 112 that identifies the resources by crawling and indexing the resources provided by the content publishers on the websites 104. Data about the resources can be indexed based on the resource to which the data corresponds. The indexed and, optionally, cached copies of the resources can be stored in an indexed cache 114.

[0037] User devices 106 can submit search queries 116 to the search system 112 over the network 102. In response, the search system 112 accesses the indexed cache 114 to identify resources that are relevant to the search query 116. The search system 112 identifies the resources in the form of search results 118 and returns the search results 118 to the user devices 106 in search results pages. A search result 118 can be data generated by the search system 112 that identifies a resource that is provided in response to a particular search query, and includes a link to the resource. In some implementations, the search results 118 include the content itself, such as a map, or an answer, such as in response to a query for a store's products, phone number, address or hours of operation. In some implementations, the content management system 110 can generate search results 118 using information (e.g., identified resources) received from the search system 112. An example search result 118 can include a webpage title, a snippet of text or a portion of an image extracted from the webpage, and the URL of the webpage. Search results pages can also include one or more slots in which other content items (e.g., ads) can be presented. In some implementations, slots on search results pages or other webpages can include content slots for content items that have been provided as part of a reservation process. In a reservation process, a publisher and a content item sponsor enter into an agreement where the publisher agrees to publish a given content item (or campaign) in accordance with a schedule (e.g., provide 1000 impressions by date X) or other publication criteria. In some implementations, content items that are selected to fill the requests for content slots can be selected based, at least in part, on priorities associated with a reservation process (e.g., based on urgency to fulfill a reservation).

[0038] When a resource 105, search results 118 and/or other content are requested by a user device 106, the content management system 110 receives a request for content. The request for content can include characteristics of the slots that are defined for the requested resource or search results page, and can be provided to the content management system 110.

[0039] For example, a reference (e.g., URL) to the resource for which the slot is defined, a size of the slot, and/or media types that are available for presentation in the slot can be provided to the content management system 110. Similarly, keywords associated with a requested resource ("resource keywords") or a search query 116 for which search results are requested can also be provided to the content management system 110 to facilitate identification of content that is relevant to the resource or search query 116.

[0040] Based at least in part on data included in the request, the content management system 110 can select content that is eligible to be provided in response to the request (“eligible content items”). For example, eligible content items can include eligible ads having characteristics matching the characteristics of ad slots and that are identified as relevant to specified resource keywords or search queries 116. In some implementations, the selection of the eligible content items can further depend on user signals, such as demographic signals and behavioral signals.

[0041] The content management system 110 can select from the eligible content items that are to be provided for presentation in slots of a resource or search results page based at least in part on results of an auction (or by some other selection process). For example, for the eligible content items, the content management system 110 can receive offers from content sponsors 108 and allocate the slots, based at least in part on the received offers (e.g., based on the highest bidders at the conclusion of the auction or based on other criteria, such as those related to satisfying open reservations). The offers represent the amounts that the content sponsors are willing to pay for presentation (or selection or other interaction with) of their content with a resource or search results page. For example, an offer can specify an amount that a content sponsor is willing to pay for each 1000 impressions (i.e., presentations) of the content item, referred to as a CPM bid. Alternatively, the offer can specify an amount that the content sponsor is willing to pay (e.g., a cost per engagement) for a selection (i.e., a click-through) of the content item or a conversion following selection of the content item. For example, the selected content item can be determined based on the offers alone, or based on the offers of each content sponsor being multiplied by one or more factors, such as quality scores derived from content performance, landing page scores, and/or other factors.

[0042] The content management system 110 can include plural engines. A spam detection engine 122 can detect spammers among users who access content provided by the content management system 110. Detecting spammers can include evaluating user interactions with the content items, e.g., using information that associates user interactions (e.g., click patterns) with spammers and non-spammers. These spammers are suspected spammers for the purposes of determining false positive ratios.

[0043] A content selection engine 121 can select and provide content that is used to detect spammers. For example, the content selection engine 121 may select regular content items 125 in the vast majority of situations, and test content items 126 to a smaller percentage of users who are suspected spammers. In some implementations,

different types of test content items 126 can be served to the same user, e.g., depending on a likelihood that the user is a spammer or a spammer category currently associated with the user. For example, a first type of content item can be provided as part of first-level testing that tests whether a potential spammer is a likely spammer, and a second different type of content item can be provided as part of second-level testing that tests whether a likely spammer is confirmed a spammer. These tests can also conclude that a user is not spammer at all. Other numbers of levels of testing can be used, e.g., in more sophisticated frameworks that identify more sub-categories of potential, likely and other spammers.

[0044] A false positive engine 123, for example, can analyze user responses by suspected spammers to content items chosen for presentation by the content selection engine 121. The false positive engine 123 can mark a potential spammer as a likely spammer depending on the potential spammer's interaction with the first type of content item. The false positive engine 123 can also mark a likely spammer as a confirmed spammer depending on the likely spammer's interaction with the second different type of content item. A data store of marked spammers 128 can store information about spammers, including, for example, identifiers of potential spammers, likely spammers, confirmed spammers, and identifications of users who have been determined to be non-spammers. In some implementations, information about a particular user can include a percentage likelihood that the user is a spammer, or a confidence level that a user is a non-spammer. Other probabilistic and/or categorical information can be stored for users identified as spammers and non-spammers.

[0045] The spam detection engine 122 and false positive engine 123 can track information over time regarding numbers of users who are or are not spammers. For example, the false positive engine 123 can determine false positive ratios of the spam detection engine 122 based at least in part on a number of times users are marked as potential spammers versus the number of times that users are ultimately marked as spammers.

[0046] In some implementations, requests for content can be handled by a request handler 124, e.g., which can determine and confirm content eligibility before the content selection engine 121 generates either a test version of content or a non-test version as will be described in greater detail below. The request handler 124 can also re-direct incoming request so that the spam detection engine 122 can detect spammers and the false

positive engine 123 can perform tests related to detecting false positive ratios and more accurately identifying user spammers (or not).

[0047] A conversion can be said to occur when a user performs a particular transaction or action related to a content item provided with a resource or search results page. What constitutes a conversion may vary from case-to-case and can be determined in a variety of ways. For example, a conversion may occur when a user clicks on a content item (e.g., an ad), is referred to a webpage, and consummates a purchase there before leaving that webpage. A conversion can also be defined by a content provider to be any measurable or observable user action, such as downloading a white paper, navigating to at least a given depth of a website, viewing at least a certain number of webpages, spending at least a predetermined amount of time on a web site or webpage, registering on a website, or experiencing media. Other actions that constitute a conversion can also be used

[0048] FIG. 2 shows an example system 200 for determining false positive ratios in spam detectors. For example, the content management system 110 can determine the false positive ratios for users using the user devices 106 when accessing content (e.g., advertisements) sponsored by content sponsors 108. An example sequence of steps follows that provides one possible series of events for providing advertisements in this way. Other forms of content can be provided (e.g., other forms of sponsored or non-sponsored content).

[0049] At step 1, the content management system 110 can identify a spam detector 202, such as the spam detection engine 122, for detecting spammers, including users of user devices 106. For example, the spam detection engine 122 can determine, based at least on interactions by a user 201, whether the user 201 is a suspected spammer.

[0050] At step 2, the content management system 110 can receive an input 204 that a particular user is a suspected spammer. For example, the spam detection engine 122 may identify the user 201 as a spammer based on user interactions of the user 201, such as reacting to content presented on the user device 106. The spammers identified in this step include spammers that are part of a false positive population, e.g., including users who were incorrectly identified as spammers. However, in the steps that follow, the system 200 may ultimately determine that the spammers identified in this step are either confirmed spammers or non-spammers.

[0051] At step 3, the content management system 110 can record an identifier 206 for the user 201, and the user 201 can be marked as a potential spammer. As an example, the

content management system 110 or the false positive engine 123 can store an IP address, a cookie, or some other identifier in the marked spammers 128 to identify the user 201 as a potential spammer. An identifier also stored for the entry in the marked spammers 128 can be associated with the user 201 or the user device 106.

[0052] In some implementations, the false positive engine 123 can classify users such as the user 201 as certain types of spammers. For example, if information is received that the user 201 is a potential spammer, then the false positive engine 123 can assign that level of spammer to the user, as opposed to marking the user as a likely spammer or confirmed spammer, which can happen in other circumstances, as described below.

[0053] At step 4, the content management system 110 can receive a request for content (e.g., a query 208) from the potential spammer, or an indication can be received the potential spammer has requested to view or interact with a content item. For example, sometime after being marked as a potential spammer, the user 201 may make a selection on a web page that generates a request for content from the content management system 110. In some implementations, requests for content, including queries, can be handled at the content management system 110 by the request handler 124.

[0054] At step 5, the content management system 110 can provide the potential spammer with a test content item 210, e.g., a first type of content item that tests whether the potential spammer is a likely spammer. For example, in a first-level test of the user, the first type of content item that is provided can be an advertisement with a changeable selection area, such as a blinking advertisement. The content management system 110 can also provide other types of test content items 126 in addition to blinking advertisements that are used in first-level tests.

[0055] At step 6a, when the potential spammer interacts with the first type of content item, the potential spammer is marked as a likely spammer at step 6b. For example, if the first type of content item is a blinking advertisement, the false positive engine 123 can determine an interaction latency associated with the user's interaction with the changeable selection area. If the user 201 makes a selection at a time when the advertisement is not visible (e.g., an action that a bot may do), then the false positive engine 123 can mark 214 the user as a likely spammer.

[0056] In some implementations, click logs 129 can be maintained that include information about the number and time of user interactions with content have occurred (such as clicks performed by users, e.g., including by potential spammers in response to presented content). For example, the false positive engine 123 can use information from

the click logs 129 to help to identify potential spammers who have a large number of repeated clicks (e.g., that may signal repetitive clicks by a bot). Further, there may be clicks by users that do not match the expected click patterns that would be expected, for example, in response by an actual human being to a blinking advertisement.

[0057] At step 7, the content management system 110 can receive a request for content or interaction with content from/by the likely spammer, or an indication can be received that the likely spammer has requested to view a content item. For example, sometime after the user 201 is marked as a likely spammer, the user 201 may make a selection on the same or a different web page, causing a request for content to be generated for receipt by the content management system 110.

[0058] At step 8, the content management system 110 can provide the likely spammer with a version of the test content item 210 that tests whether the likely spammer is a confirmed spammer. For example, the test content item 210 presented to the likely spammer can be a different type of test content item 126 than the content item presented to a potential spammer. Selection of the second different type of content item, once presented, represents a high probability that the likely spammer is a confirmed spammer. For example, the second different type of content can be part of a second-level test for determining whether a likely spammer is a confirmed spammer. Types of content that can accomplish this include a junk advertisement, an empty advertisement, a hidden advertisement, an advertisement that is in a different language than the language associated with the user, or a mismatched advertisement (e.g., the advertisement's image does not match the text).

[0059] At step 9a, when the likely spammer interacts with the second different type of content item, the likely spammer is marked as a confirmed spammer at step 9b. For example, if the second different type of content is junk advertisement for which a user interaction 216 is consistent with actions of a non-human, then the false positive engine 123 can mark the user as confirmed spammer 218.

[0060] At step 10, the content management system 110 can determine a false positive ratio 220 of the spam detector based at least in part on a number of times that users are marked as potential spammers versus the number of times that users are ultimately marked as spammers. In some implementations, determining a false positive ratio can use the following formula:

$$\text{False Positive Ratio} = \text{Non-Spammer Clicks} / (\text{Non-Spammer Clicks} + \text{Spammer Clicks}) \quad (1)$$

[0061] For at least the users who are known to be non-spammers, the content selection engine 121 can select conventional content items 125 to be provided in response to a request for content. For some or all of the users who may be spammers, the content selection engine 121 can instead select test or non-paying content items 126. In some implementations, e.g., when determining false positive ratios, the content selection engine 121 can use test content items 126 for a representative sample so that test content items need not be provided to an entire population.

[0062] FIG. 3 is a block diagram showing example characteristics 300 of different levels of tests 302-306 performed over time to detect spammers. The tests 302-306, for example, can be provided to users at different levels, e.g., the first level being when a user is considered to be a potential spammer. As time progresses, and as the user completes more tests, subsequent tests 204-206 can change over time.

[0063] The characteristics 300, for example, indicate that early tests are done when the user has a high reputation 300a (e.g., may not yet be considered a potential spammer). Also, in early-level tests, a quality of service 300b can be high, but test efficiency 300c may be low. For example, test 302 (e.g., a test using a low click-through ratio) can be associated with a user having a high reputation 300a. Further, the test 302 can provide a high quality of service 300b but a low test efficiency 300c. At test 304 (e.g., a test using an interstitial page), the user's reputation 300a may be high and resultant test may not substantially affect the quality of service 300b. In this example, test efficiency 300c can be in the medium range. Finally, by test 306 (e.g., a test using junk ads), the user's reputation 300a may be low (e.g., already determined to be a likely spammer), the quality of service 300b may be low (e.g., more aggressive-test content is provided), e.g., resulting in a high test efficiency 300c.

[0064] In the early-level tests (e.g., test 302), it may be more likely that a test creates a false positive 308. However, as time goes on, the later-level tests (e.g., test 306) may be less likely to produce false positives 308 and more likely to result in producing true positives 310 (e.g., the user is a confirmed spammer) or determining that the user is a non-spammer.

[0065] FIG. 4 is a flowchart of an example process 400 for determine a false positive ratio of a spam detector. In some implementations, the content management system 110 can perform steps of the process 400 using instructions that are executed by one or more

processors. FIGS. 1-3 are used to provide example structures for performing the steps of the process 400.

[0066] A spam detector is identified that is operable to determine when an interaction with a content item by a user is true or is more likely to be spam (402). For example, the content management system 110 can include the spam detection engine 122.

[0067] An input is received that a particular user is a spammer (404). For example, the spam detection engine 122 can identify the user 201 as a potential spammer, e.g., based on user interactions with content presented on the client device 106. Users identified in this way include users that may be falsely marked as spammers.

[0068] An identifier is recorded for the user, and the user is marked as a potential spammer (406). For example, the user classification can store an entry in the marked spammers 128 that identifies the user 201 as a potential spammer. The entry can include an identifier such as the IP address of the client device 106, the identifier of a cookie, or some other identifier.

[0069] A request for content (e.g., a query or interaction) is received from a potential spammer (408). As an example, the content management system 110 can receive a request for content from the user device 106, such as to provide an advertisement to fill an advertisement slot on a web page viewed by the user 201.

[0070] The potential spammer is provided with a first type of content item that tests whether the potential spammer is a likely spammer (410). For example, the content selection engine 121 can select an interstitial advertisement from the test content items 126 or some other type of first content item designed for determining whether a potential spammer can be identified as a likely spammer.

[0071] Another type of content item that can be provided in this case is a blinking advertisement that includes at least first and second states. For example, the first state can be a state in which the blinking advertisement is visible to the user, and the second state can be a state in which the blinking advertisement is not visible to the user. Tests using blinking advertisements, for example, can determine whether a user is a spammer by correlating a click event time with a blinking period, e.g., between first and second states when the advertisement alternates between being visible and invisible.

[0072] In some implementations, the first type of content item can be a content item that includes a changeable selection area, and the process 400 can further include determining an interaction latency for the user's interaction with the changeable selection area. For example, this type of content item can detect some bots that may repeatedly

click until they succeed. A user can be determined to be a spammer in these types of tests because click latencies can signal a difference between bots and actual human beings in solving the problem of where to click.

[0073] In some implementations, test advertisements or other content can include the use for challenge-response tests (e.g., CAPTCHA tests). For example, these tests can help determine when responses are provided by a user versus a machine (e.g., computer).

[0074] When the potential spammer interacts with the first type of content item, the potential spammer is marked as a likely spammer (412). For example, depending on how the user 201 interacts with an interstitial, blinking or other type of advertisement that is provided as a first-level content item, the false positive engine 123 can mark the user 201 as a likely spammer in the marked spammers 128.

[0075] A subsequent request for content is received from the likely spammer (414). For example, the user 201 may issue another query or perform some other action on the user device 106 that causes the user device 106 to send a request for content to the content management system 110.

[0076] The likely spammer is provided with a second different type of content item that tests whether the likely spammer is a spammer (416). For example, the second different type of content item can be a junk advertisement, an empty advertisement, an advertisement that is in a different language than that associated with the user, or a mismatched advertisement. Selection of the second different type of content item, once presented, represents a high probability that the likely spammer is a spammer. For example, by looking up the user in the marked spammers 128, the false positive engine 123 can learn that the user 201 is a potential spammer. As a result, the content selection engine 121 can select a second-level advertisement from the test content items 126.

[0077] In some implementations, the tests associated with the first and second different types of content items can use different probabilities. For example, depending on the test, the probability that a user is confidently determined to be a potential spammer, spammer (or a non-spammer) can vary.

[0078] When the likely spammer interacts with the second different type of content item, the likely spammer is marked as a spammer (418). For example, if the user 201 is presented with an empty advertisement or some other advertisement with which actual users typically do not interact, then the user is marked as a confirmed spammer.

[0079] A false positive ratio of the spam detector is determined based at least in part on a number of times that users are marked as potential spammers versus the number of

times that users are ultimately marked as confirmed spammers (420). For example, the false positive engine 123 can calculate an accurate false positive ratio based on numbers of spammer clicks and non-spammer clicks (e.g., using formula 1 above). Other ways can be used, such as determining a ratio of users determined to be non-spammers due to not being marked as spammers and a sum of the users determined to be non-spammers and the users determined to be spammers.

[0080] In some implementations, the process 400 can further include adjusting the spam detector based at least in part on the determined false positive ratio. For example, based on information learned over time about which users are ultimately determined to be spammers or non-spammers, the techniques used by the spam detector 122 and the false positive engine 123 can be adjusted over time. Further, the types of test content items 125 selected by the content selection engine 121 can change, e.g., if it is determined that certain types of advertisements are more productive and/or accurate in identifying spammers.

[0081] In some implementations, the process 400 can further include repeating the method after a predetermined time or when a predetermined condition is met in order to determine a new false positive ratio. For example, the content management system 110 may initially calculate a false positive ratio at one time, and over time re-calculate the false positive ratio when additional information (e.g., including user interactions, click rates, etc.) is available.

[0082] In some implementations, the process 400 can further include adjusting a click-through amount for an advertiser for a given campaign based at least in part on the false positive ratio. For example, an advertiser's click-through amount can be based on determining spammers using the techniques described in this document, e.g., that might be based on a significant number of false positives. However, the false positive ratio calculated herein can be used to adjust the advertiser's click-through amount, e.g., as a product of the false positive ratio.

[0083] In some implementations, instead of having two tests, e.g., a level-one test involving a first type of content and a level-two test involving a second type of content, a single test can be used. For example, in click-to-call situations, a potential spammer can be directed to a different location than a non-spammer. The location may be, for example, a different phone number, or a regular phone number with an accompanying indicator that the incoming call is potentially from a spammer. In some implementations, level-one and level-two tests can be combined in some way, e.g., so testing for a potential

or likely spammer can be accomplished with a single type of content item (e.g., a click-to-call advertisement).

[0084] In some implementations, the first type of content item can be a content item that includes a fixed user-visible selection area, and the process 400 can further include identifying an interaction location for where an entity interacted with the fixed selection area so as to be able to infer whether the interaction was associated with a human user or a machine. The process 400 can further include determining interaction locations that are likely associated with human interaction and interaction locations that are unlikely associated with human interactions.

[0085] In some implementations, the process 400 can further include dividing users into multiple groups, providing first and/or second types of content items to users with different probabilities based on their respective groups, and using statistical analysis of false positives of the multiple groups to determine spamminess of entire groups.

[0086] In some implementations, the first type of content can include a low click-through ratio advertisement. Upon detecting an interaction with the low click-through ratio advertisement, the user can be labeled a potential spammer.

[0087] FIG. 5 is a flowchart of an example process 500 for classifying users as spammers. In some implementations, the content management system 110 can perform steps of the process 500 using instructions that are executed by one or more processors. FIGS. 1-3 are used to provide example structures for performing the steps of the process 500.

[0088] An indication is received that an interaction associated with a user is determined to be spam, where spam represents a false interaction with a content item (502). For example, the spam detection engine 122 can identify the user 201 as a potential spammer.

[0089] An identifier associated with the user is saved and the user is marked as a potential spammer (504). As an example, the false positive engine 123 can mark the user 201 as a potential spammer, and the information can be stored in the marked spammers 128 along with an identifier (e.g., IP address or cookie).

[0090] A request for content is received from the potential spammer (506). As an example, the content management system 110 can receive a request for content from the user device 106, such as to provide an advertisement to fill an advertisement slot on a web page viewed by the user 201.

[0091] The potential spammer is provided with a first type of content item in response to the request or in lieu of directing a user to a landing page associated with the advertisement that tests whether the potential spammer is a likely spammer (508). For example, the content selection engine 121 can select an interstitial advertisement from the test content items 126 or some other type of first content item (e.g., a blinking advertisement) designed for determining whether the user 201 can be identified as a likely spammer. In another example, the user 201 can be directed to a different landing page. In yet another example, the first type of content can be click-to-call content, e.g., that provides a different phone number or outcome subsequent to the click.

[0092] When the potential spammer interacts with the first type of content item, the potential spammer is marked as a likely spammer (510). As an example, depending on how the user 201 interacts with an interstitial, blinking or other type of advertisement that is provided as a first-level content item, the false positive engine 123 can mark the user 201 as a likely spammer in the marked spammers 128.

[0093] Thereafter, a subsequent request for content is received from the likely spammer (512). For example, the user 201 may issue another query or perform some other action on the user device 106 that causes the user device 106 to send a request for content to the content management system 110.

[0094] The likely spammer is provided with a second different type of content item in response to the request (514). The second different type of content item is one that tests whether the likely spammer is a confirmed spammer, wherein interaction with or selection of the second different type of content item, once presented, represents a high probability that the likely spammer is a spammer. For example, the second different type of content item can be a junk advertisement, an empty advertisement, an advertisement that is in a different language than that associated with the user, or a mismatched advertisement. In another example, the second different type of content can be the click-to-call details that follow the selection of click-to-call content (e.g., a first type of content).

[0095] When the likely spammer interacts with the second different type of content item, the likely spammer is marked as a spammer (516). As an example, if the user 201 is presented with an empty advertisement or some other advertisement with which actual users (e.g., non-spammers) typically do not interact, then the user 201 can be marked as a confirmed spammer. In the example of click-to-call content, the classification of the user

201 (e.g., as a confirmed spammer) can be based on the call that occurs or other user actions, such as message interaction and/or conversation.

[0096] A final determination is stored for the user based at least in part on the interaction with the second different type of content item by the user (518). For example, the false positive detector 122 can mark the user 201 as a confirmed spammer in the marked spammers 128.

[0097] In some implementations, the second type of content used in either or both of processes 400 and 500 can include a user login requirement, an email acknowledgment, or some other user authentication. In some implementations, this type of action can produce some kind of user identifier, which can also be linked, for example, with a call duration or some other user interaction or event.

[0098] In some implementations, the techniques for detecting spammers and determining false positive ratios can be used in a real-time, distributed spam filtering system. For example, one or more content management systems 110 (including spam detection engines 122 and false positive engines 123) can exist at each of one or more data centers that provide data service to a network (or group of interconnected networks such as the Internet). Further, the data centers may also provide content such as advertisements. Here, spam classification can occur in real time. Experiments and projections have shown that this can result in providing, for example, spam detection coverage on about 99.99% of Internet traffic within a deadline 10ms when the traffic rate is below a threshold of 10K QPS. Advertisement servers, including content management systems 110, can take advantage of this capacity, providing inline spam/spammer classification and showing different types of content items (e.g., advertisements and/or providing different landing pages) for different types of users (e.g., spammer and others).

[0099] In some implementations, a small subset of available spam filters can be used and still provide adequate protection from spam and spammers, e.g., instead of using the full set of available filters. For example, experiments and projections have shown that the use of approximately five of the most effective spam filters (e.g., including both stateless and stateful filters) can filter more than 75% of spam traffic in real time. Further, real-time filters can be selected and configured to ensure an inclusion property (e.g., eliminating over-filtering of content as compared with the online and offline filters). Some examples of filters (e.g., in one possible descending order of effectiveness as determined from experiments) include following filters. In a near duplicate filter, for example, for each impression, if there are two clicks close together in time and both

clicks are registered on the same impression, then the first click can be marked as spam. In a duplicate filter, for example, for each query, if there are two clicks close together in time, then each click can be marked as spam. Other filters include a persistent IP-to-DP filter, a bad web property filter, an expired filter, an unexpected user filter, a click delay by Wp filter, a manual blacklisted filter, a blacklisted filter, a bad user agent filter, and a CPA CPP by query filter.

[00100] In some implementations, load balancing can be used between the one or more data centers nodes. For example, the capacity of each filter may be different, and each data center may house just one or a few of the filters.

[00101] FIG. 6 is a block diagram of example computing devices 600, 650 that may be used to implement the systems and methods described in this document, as either a client or as a server or plurality of servers. Computing device 600 is intended to represent various forms of digital computers, such as laptops, desktops, workstations, personal digital assistants, servers, blade servers, mainframes, and other appropriate computers. Computing device 600 is further intended to represent any other typically non-mobile devices, such as televisions or other electronic devices with one or more processors embedded therein or attached thereto. Computing device 650 is intended to represent various forms of mobile devices, such as personal digital assistants, cellular telephones, smartphones, and other computing devices. The components shown here, their connections and relationships, and their functions, are meant to be exemplary only, and are not meant to limit implementations of the inventions described and/or claimed in this document.

[00102] Computing device 600 includes a processor 602, memory 604, a storage device 606, a high-speed interface 608 connecting to memory 604 and high-speed expansion ports 610, and a low speed interface 612 connecting to low speed bus 614 and storage device 606. Each of the components 602, 604, 606, 608, 610, and 612, are interconnected using various busses, and may be mounted on a common motherboard or in other manners as appropriate. The processor 602 can process instructions for execution within the computing device 600, including instructions stored in the memory 604 or on the storage device 606 to display graphical information for a GUI on an external input/output device, such as display 616 coupled to high speed interface 608. In other implementations, multiple processors and/or multiple buses may be used, as appropriate, along with multiple memories and types of memory. Also, multiple computing devices

600 may be connected, with each device providing portions of the necessary operations (e.g., as a server bank, a group of blade servers, or a multi-processor system).

[00103] The memory 604 stores information within the computing device 600. In one implementation, the memory 604 is a computer-readable medium. In one implementation, the memory 604 is a volatile memory unit or units. In another implementation, the memory 604 is a non-volatile memory unit or units.

[00104] The storage device 606 is capable of providing mass storage for the computing device 600. In one implementation, the storage device 606 is a computer-readable medium. In various different implementations, the storage device 606 may be a floppy disk device, a hard disk device, an optical disk device, or a tape device, a flash memory or other similar solid state memory device, or an array of devices, including devices in a storage area network or other configurations. In one implementation, a computer program product is tangibly embodied in an information carrier. The computer program product contains instructions that, when executed, perform one or more methods, such as those described above. The information carrier is a computer- or machine-readable medium, such as the memory 604, the storage device 606, or memory on processor 602.

[00105] The high speed controller 608 manages bandwidth-intensive operations for the computing device 600, while the low speed controller 612 manages lower bandwidth-intensive operations. Such allocation of duties is exemplary only. In one implementation, the high-speed controller 608 is coupled to memory 604, display 616 (e.g., through a graphics processor or accelerator), and to high-speed expansion ports 610, which may accept various expansion cards (not shown). In the implementation, low-speed controller 612 is coupled to storage device 606 and low-speed expansion port 614. The low-speed expansion port, which may include various communication ports (e.g., USB, Bluetooth, Ethernet, wireless Ethernet) may be coupled to one or more input/output devices, such as a keyboard, a pointing device, a scanner, or a networking device such as a switch or router, e.g., through a network adapter.

[00106] The computing device 600 may be implemented in a number of different forms, as shown in the figure. For example, it may be implemented as a standard server 620, or multiple times in a group of such servers. It may also be implemented as part of a rack server system 624. In addition, it may be implemented in a personal computer such as a laptop computer 622. Alternatively, components from computing device 600 may be combined with other components in a mobile device (not shown), such as device 650. Each of such devices may contain one or more of computing device 600, 650, and an

entire system may be made up of multiple computing devices 600, 650 communicating with each other.

[00107] Computing device 650 includes a processor 652, memory 664, an input/output device such as a display 654, a communication interface 666, and a transceiver 668, among other components. The device 650 may also be provided with a storage device, such as a microdrive or other device, to provide additional storage. Each of the components 650, 652, 664, 654, 666, and 668, are interconnected using various buses, and several of the components may be mounted on a common motherboard or in other manners as appropriate.

[00108] The processor 652 can process instructions for execution within the computing device 650, including instructions stored in the memory 664. The processor may also include separate analog and digital processors. The processor may provide, for example, for coordination of the other components of the device 650, such as control of user interfaces, applications run by device 650, and wireless communication by device 650.

[00109] Processor 652 may communicate with a user through control interface 658 and display interface 656 coupled to a display 654. The display 654 may be, for example, a TFT LCD display or an OLED display, or other appropriate display technology. The display interface 656 may comprise appropriate circuitry for driving the display 654 to present graphical and other information to a user. The control interface 658 may receive commands from a user and convert them for submission to the processor 652. In addition, an external interface 662 may be provided in communication with processor 652, so as to enable near area communication of device 650 with other devices. External interface 662 may provide, for example, for wired communication (e.g., via a docking procedure) or for wireless communication (e.g., via Bluetooth or other such technologies).

[00110] The memory 664 stores information within the computing device 650. In one implementation, the memory 664 is a computer-readable medium. In one implementation, the memory 664 is a volatile memory unit or units. In another implementation, the memory 664 is a non-volatile memory unit or units. Expansion memory 674 may also be provided and connected to device 650 through expansion interface 672, which may include, for example, a subscriber identification module (SIM) card interface. Such expansion memory 674 may provide extra storage space for device 650, or may also store applications or other information for device 650. Specifically, expansion memory 674 may include instructions to carry out or supplement the processes

described above, and may include secure information also. Thus, for example, expansion memory 674 may be provide as a security module for device 650, and may be programmed with instructions that permit secure use of device 650. In addition, secure applications may be provided via the SIM cards, along with additional information, such as placing identifying information on the SIM card in a non-hackable manner.

[00111] The memory may include for example, flash memory and/or MRAM memory, as discussed below. In one implementation, a computer program product is tangibly embodied in an information carrier. The computer program product contains instructions that, when executed, perform one or more methods, such as those described above. The information carrier is a computer- or machine-readable medium, such as the memory 664, expansion memory 674, or memory on processor 652.

[00112] Device 650 may communicate wirelessly through communication interface 666, which may include digital signal processing circuitry where necessary. Communication interface 666 may provide for communications under various modes or protocols, such as GSM voice calls, SMS, EMS, or MMS messaging, CDMA, TDMA, PDC, WCDMA, CDMA2000, or GPRS, among others. Such communication may occur, for example, through radio-frequency transceiver 668. In addition, short-range communication may occur, such as using a Bluetooth, WiFi, or other such transceiver (not shown). In addition, GPS receiver module 670 may provide additional wireless data to device 650, which may be used as appropriate by applications running on device 650.

[00113] Device 650 may also communicate audibly using audio codec 660, which may receive spoken information from a user and convert it to usable digital information. Audio codec 660 may likewise generate audible sound for a user, such as through a speaker, e.g., in a handset of device 650. Such sound may include sound from voice telephone calls, may include recorded sound (e.g., voice messages, music files, etc.) and may also include sound generated by applications operating on device 650.

[00114] The computing device 650 may be implemented in a number of different forms, as shown in the figure. For example, it may be implemented as a cellular telephone 680. It may also be implemented as part of a smartphone 682, personal digital assistant, or other mobile device.

[00115] Various implementations of the systems and techniques described here can be realized in digital electronic circuitry, integrated circuitry, specially designed ASICs (application specific integrated circuits), computer hardware, firmware, software, and/or combinations thereof. These various implementations can include implementation in one

or more computer programs that are executable and/or interpretable on a programmable system including at least one programmable processor, which may be special or general purpose, coupled to receive data and instructions from, and to transmit data and instructions to, a storage system, at least one input device, and at least one output device.

[00116] These computer programs (also known as programs, software, software applications or code) include machine instructions for a programmable processor, and can be implemented in a high-level procedural and/or object-oriented programming language, and/or in assembly/machine language. Other programming paradigms can be used, e.g., functional programming, logical programming, or other programming. As used herein, the terms “machine-readable medium” “computer-readable medium” refers to any computer program product, apparatus and/or device (e.g., magnetic discs, optical disks, memory, Programmable Logic Devices (PLDs)) used to provide machine instructions and/or data to a programmable processor, including a machine-readable medium that receives machine instructions as a machine-readable signal. The term “machine-readable signal” refers to any signal used to provide machine instructions and/or data to a programmable processor.

[00117] To provide for interaction with a user, the systems and techniques described here can be implemented on a computer having a display device (e.g., a CRT (cathode ray tube) or LCD (liquid crystal display) monitor) for displaying information to the user and a keyboard and a pointing device (e.g., a mouse or a trackball) by which the user can provide input to the computer. Other kinds of devices can be used to provide for interaction with a user as well; for example, feedback provided to the user can be any form of sensory feedback (e.g., visual feedback, auditory feedback, or tactile feedback); and input from the user can be received in any form, including acoustic, speech, or tactile input.

[00118] The systems and techniques described here can be implemented in a computing system that includes a back end component (e.g., as a data server), or that includes a middleware component (e.g., an application server), or that includes a front end component (e.g., a client computer having a graphical user interface or a Web browser through which a user can interact with an implementation of the systems and techniques described here), or any combination of such back end, middleware, or front end components. The components of the system can be interconnected by any form or medium of digital data communication (e.g., a communication network). Examples of

communication networks include a local area network (“LAN”), a wide area network (“WAN”), and the Internet.

[00119] The computing system can include clients and servers. A client and server are generally remote from each other and typically interact through a communication network. The relationship of client and server arises by virtue of computer programs running on the respective computers and having a client-server relationship to each other.

[00120] While this specification contains many specific implementation details, these should not be construed as limitations on the scope of any inventions or of what may be claimed, but rather as descriptions of features specific to particular implementations of particular inventions. Certain features that are described in this specification in the context of separate implementations can also be implemented in combination in a single implementation. Conversely, various features that are described in the context of a single implementation can also be implemented in multiple implementations separately or in any suitable subcombination. Moreover, although features may be described above as acting in certain combinations and even initially claimed as such, one or more features from a claimed combination can in some cases be excised from the combination, and the claimed combination may be directed to a subcombination or variation of a subcombination.

[00121] Similarly, while operations are depicted in the drawings in a particular order, this should not be understood as requiring that such operations be performed in the particular order shown or in sequential order, or that all illustrated operations be performed, to achieve desirable results. In certain circumstances, multitasking and parallel processing may be advantageous. Moreover, the separation of various system components in the implementations described above should not be understood as requiring such separation in all implementations, and it should be understood that the described program components and systems can generally be integrated together in a single software product or packaged into multiple software products.

[00122] Thus, particular implementations of the subject matter have been described. Other implementations are within the scope of the following claims. In some cases, the actions recited in the claims can be performed in a different order and still achieve desirable results. In addition, the processes depicted in the accompanying figures do not necessarily require the particular order shown, or sequential order, to achieve desirable results. In certain implementations, multitasking and parallel processing may be advantageous.

CLAIMS

1. A computer-implemented method comprising:
 - identifying a spam detector, the spam detector operable to determine when an interaction with a content item by a user is true or is more likely to be spam;
 - receiving an input that a particular user is a spammer;
 - recording an identifier for the user and marking the user as a potential spammer;
 - receiving a request for content from a potential spammer;
 - providing the potential spammer with a first type of content item that tests whether the potential spammer is a likely spammer;
 - when the potential spammer interacts with the first type of content item, marking the potential spammer as a likely spammer;
 - receiving a subsequent request for content from the likely spammer;
 - providing the likely spammer with a second different type of content item that tests whether the likely spammer is a spammer, wherein selection of the second different type of content item, once presented, represents a high probability that the likely spammer is a spammer;
 - when the likely spammer interacts with the second different type of content item, marking the likely spammer as a spammer; and
 - determining a false positive ratio of the spam detector based at least in part on a number of times users are marked as potential spammers versus the number of times that users are ultimately marked as spammers.
2. The method of claim 1 wherein the first type of content item is an interstitial advertisement.
3. The method of claim 1 wherein the tests associated with the first and second different types of content items use different probabilities.
4. The method of claim 1 wherein the first type of content item is a blinking advertisement including at least first and second states, wherein the first state is a state in which the blinking advertisement is visible to the user, and wherein the second state is a state in which the blinking advertisement is not visible to the user.

5. The method of claim 1 wherein the first type of content item is a hidden advertisement that does not degrade a user experience.
6. The method of claim 1 wherein the first type of content item is a content item that includes a changeable selection area, and wherein the method further includes determining an interaction latency when interacting with the changeable selection area.
7. The method of claim 1 wherein the first type of content item is a content item that includes a fixed user-visible selection area, and wherein the method further includes identifying an interaction location for where an entity interacted with the fixed selection area so as to be able to infer whether the interaction was associated with a human user or a machine, wherein the method further includes determining interaction locations that are likely associated with human interaction and interaction locations that are unlikely associated with human interactions.
8. The method of claim 1 wherein the second different type of content item is selected from the group comprising a junk advertisement, an empty advertisement, a hidden advertisement, an advertisement that is in a different language than that associated with the user, or a mismatched advertisement.
9. The method of claim 1 wherein receiving an input that a particular user is a spammer includes receiving input from the spam detector that the spam detector has determined that the user is a spammer.
10. The method of claim 1 wherein determining a false positive ratio further includes determining a ratio of users determined to be non-spammers due to not being marked as spammers and a sum of the users determined to be non-spammers and the users determined to be spammers.
11. The method of claim 1 further comprising dividing users into multiple groups, providing first and/or second types of content items to users with different probabilities based on their respective groups, and using statistical analysis of false positives of the multiple groups to determine spamminess of entire groups.

12. The method of claim 1 further comprising adjusting the spam detector based at least in part on the determined false positive ratio.
13. The method of claim 1 further comprising repeating the method after a predetermined time or when a predetermined condition is met in order to determine a new false positive ratio.
14. The method of claim 1 further comprising adjusting a click-through amount for an advertiser for a given campaign based at least in part on the false positive ratio.
15. The method of claim 1 wherein the first type of content includes a low click-through ratio advertisement, and upon detecting an interaction with the low click-through ratio advertisement, labeling the user a potential spammer.
16. A computer-implemented method comprising:
 - receiving an indication that an interaction associated with a user is determined to be spam, where spam represents a false interaction with a content item;
 - saving an identifier associated with the user and marking the user as a potential spammer;
 - receiving a request for content from the potential spammer;
 - providing the potential spammer with a first type of content item in response to the request or in lieu of directing a user to a landing page associated with the advertisement that tests whether the potential spammer is a likely spammer;
 - when the potential spammer interacts with the first type of content item, marking the potential spammer as a likely spammer;
 - thereafter, receiving a subsequent request for content from the likely spammer;
 - providing the likely spammer with a second different type of content item in response to the request or in lieu of directing a user to a landing page associated with the second advertisement, the second different type of content item being one that tests whether the likely spammer is a spammer, wherein selection of the second different type of content item, once presented, represents a high probability that the likely spammer is a spammer;

when the likely spammer interacts with the second different type of content item, marking the likely spammer as a spammer; and
storing a final determination for the user based at least in part on the interaction with the second different type of content item by the user.

17. A computer program product embodied in a tangible medium including instructions, that when executed, cause one or more processors to:
- identify a spam detector, the spam detector operable to determine when an interaction with a content item by a user is true or is more likely to be spam;
 - receive an input that a particular user is a spammer;
 - record an identifier for the user and marking the user as a potential spammer;
 - receive a request for content from a potential spammer;
 - provide the potential spammer with a first type of content item that tests whether the potential spammer is a likely spammer;
 - when the potential spammer interacts with the first type of content item, mark the potential spammer as a likely spammer;
 - receive a subsequent request for content from the likely spammer;
 - provide the likely spammer with a second different type of content item that tests whether the likely spammer is a spammer, wherein selection of the second different type of content item, once presented, represents a high probability that the likely spammer is a spammer;
 - when the likely spammer interacts with the second different type of content item, mark the likely spammer as a spammer; and
 - determine a false positive ratio of the spam detector based at least in part on a number of times users are marked as potential spammers versus the number of times that users are ultimately marked as spammers.

18. The computer program product of claim 17 wherein the first type of content item is a hidden advertisement that does not degrade a user experience

19. A system comprising:
one or more processors; and

one or more memory elements including instructions that when executed cause the one or more processors to:

identify a spam detector, the spam detector operable to determine when an interaction with a content item by a user is true or is more likely to be spam;

receive an input that a particular user is a spammer;

record an identifier for the user and marking the user as a potential spammer;

receive a request for content from a potential spammer;

provide the potential spammer with a first type of content item that tests whether the potential spammer is a likely spammer;

when the potential spammer interacts with the first type of content item, mark the potential spammer as a likely spammer;

receive a subsequent request for content from the likely spammer;

provide the likely spammer with a second different type of content item that tests whether the likely spammer is a spammer, wherein selection of the second different type of content item, once presented, represents a high probability that the likely spammer is a spammer;

when the likely spammer interacts with the second different type of content item, mark the likely spammer as a spammer; and

determine a false positive ratio of the spam detector based at least in part on a number of times users are marked as potential spammers versus the number of times that users are ultimately marked as spammers.

20. The system of claim 19 wherein the first type of content item is a hidden advertisement that does not degrade a user experience.

1/6

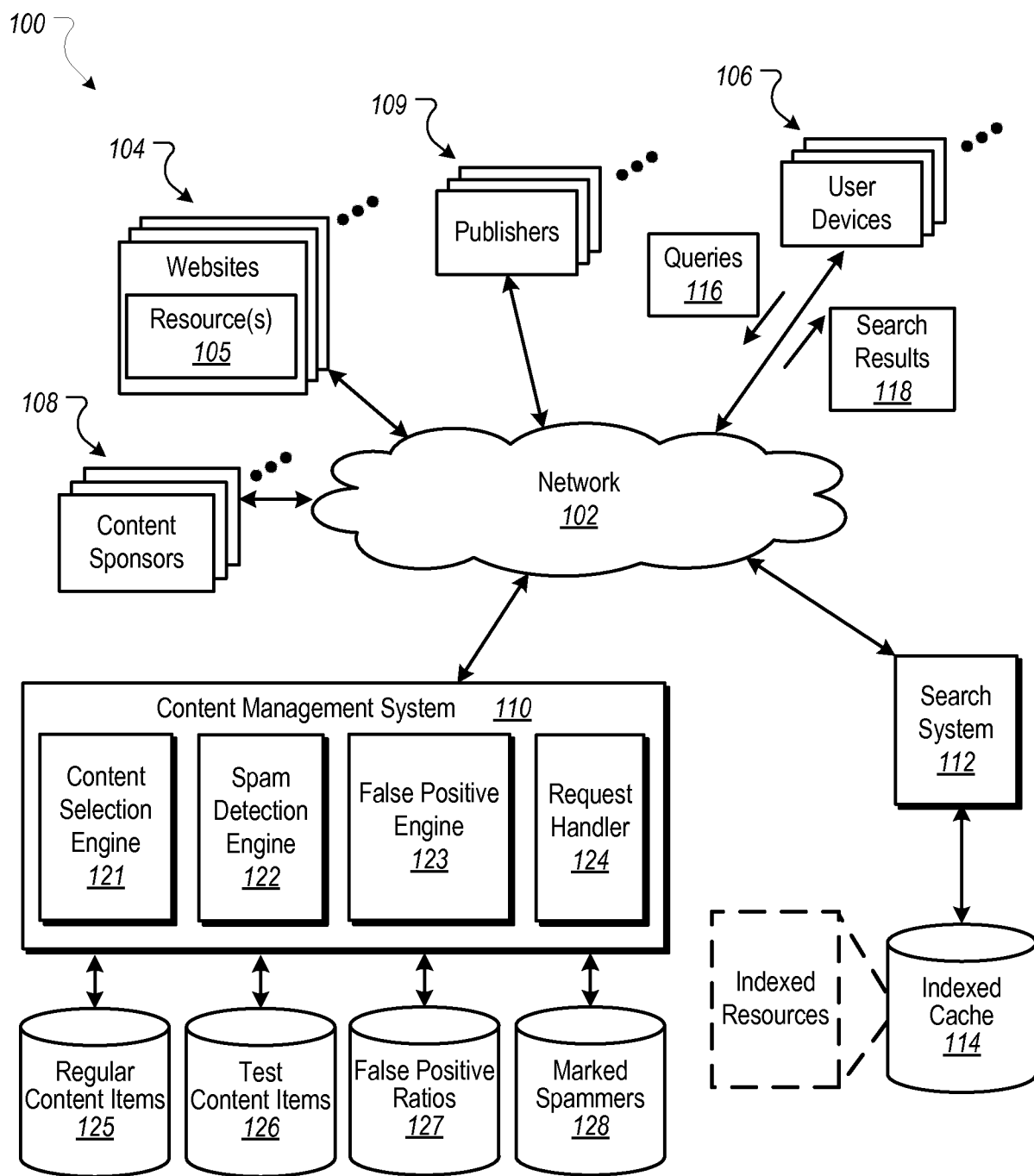
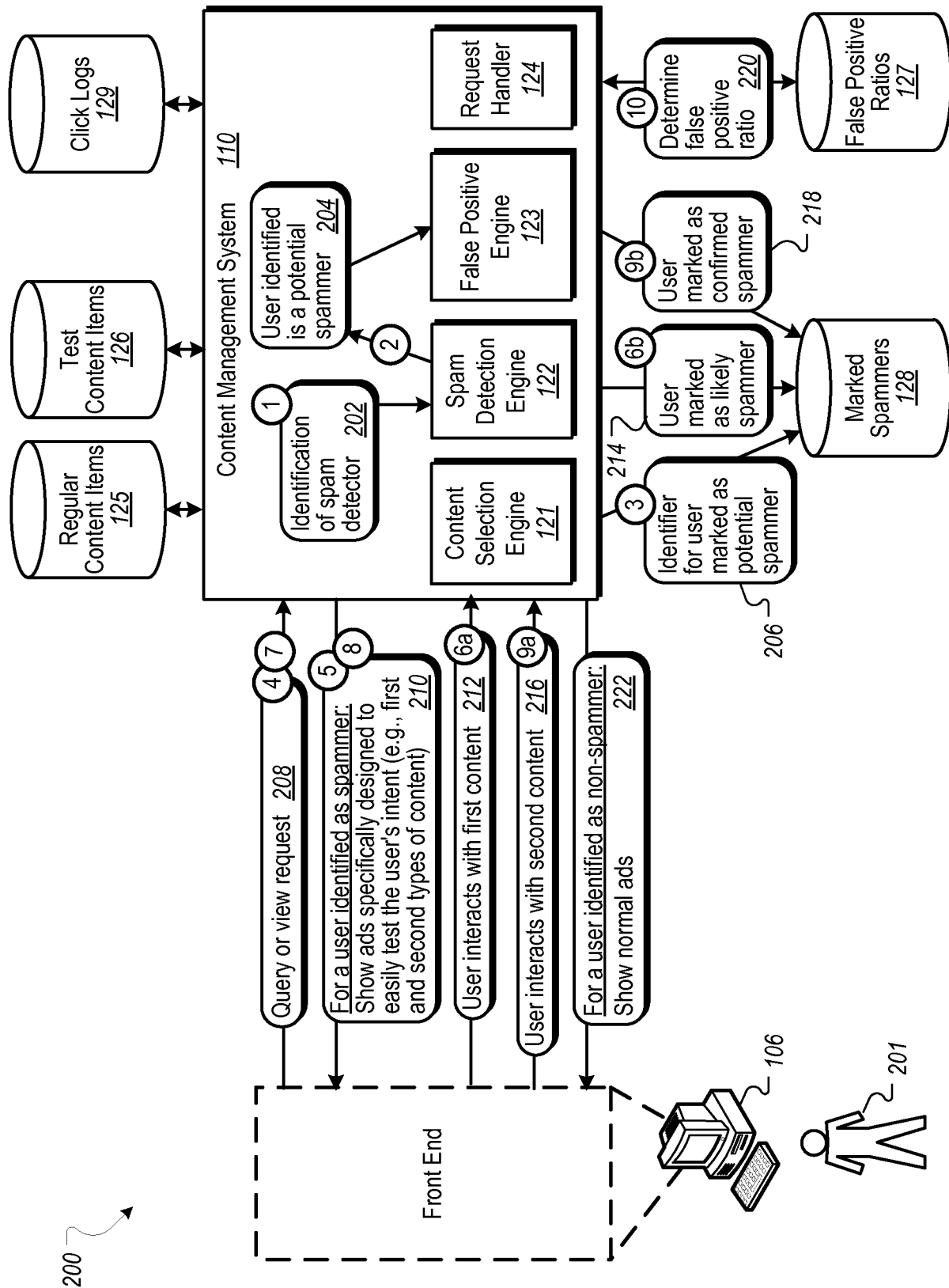


FIG. 1



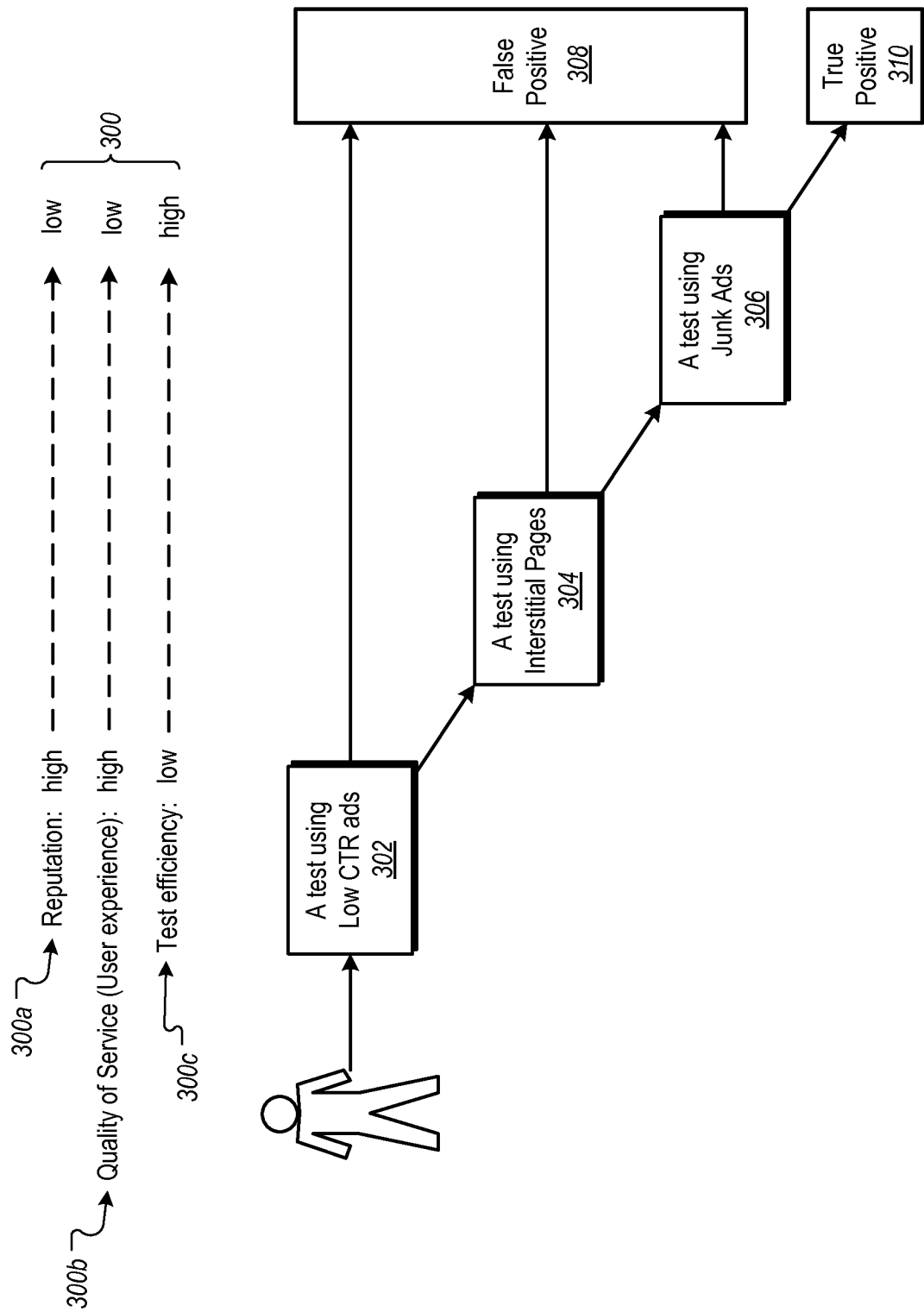


FIG. 3

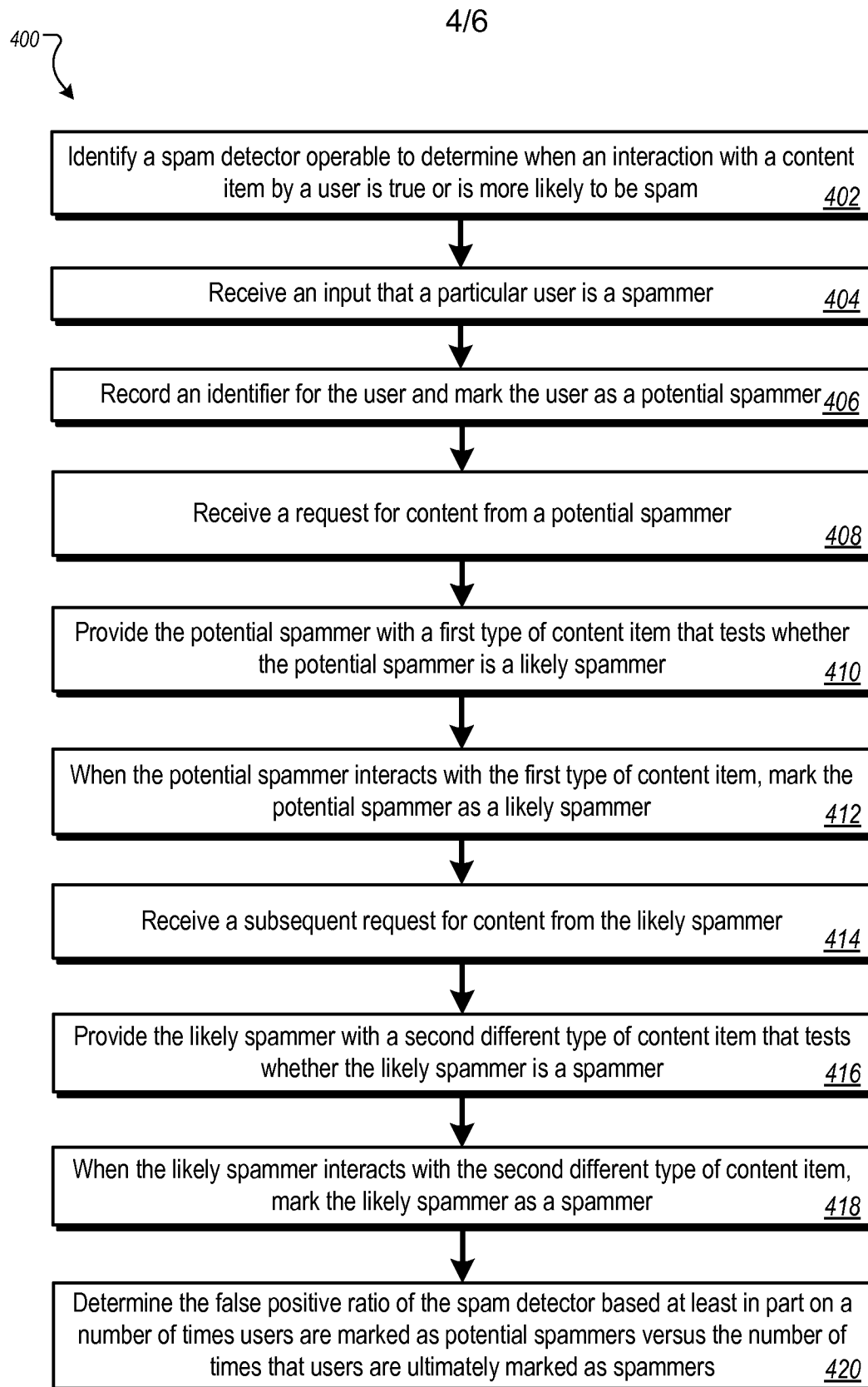


FIG. 4

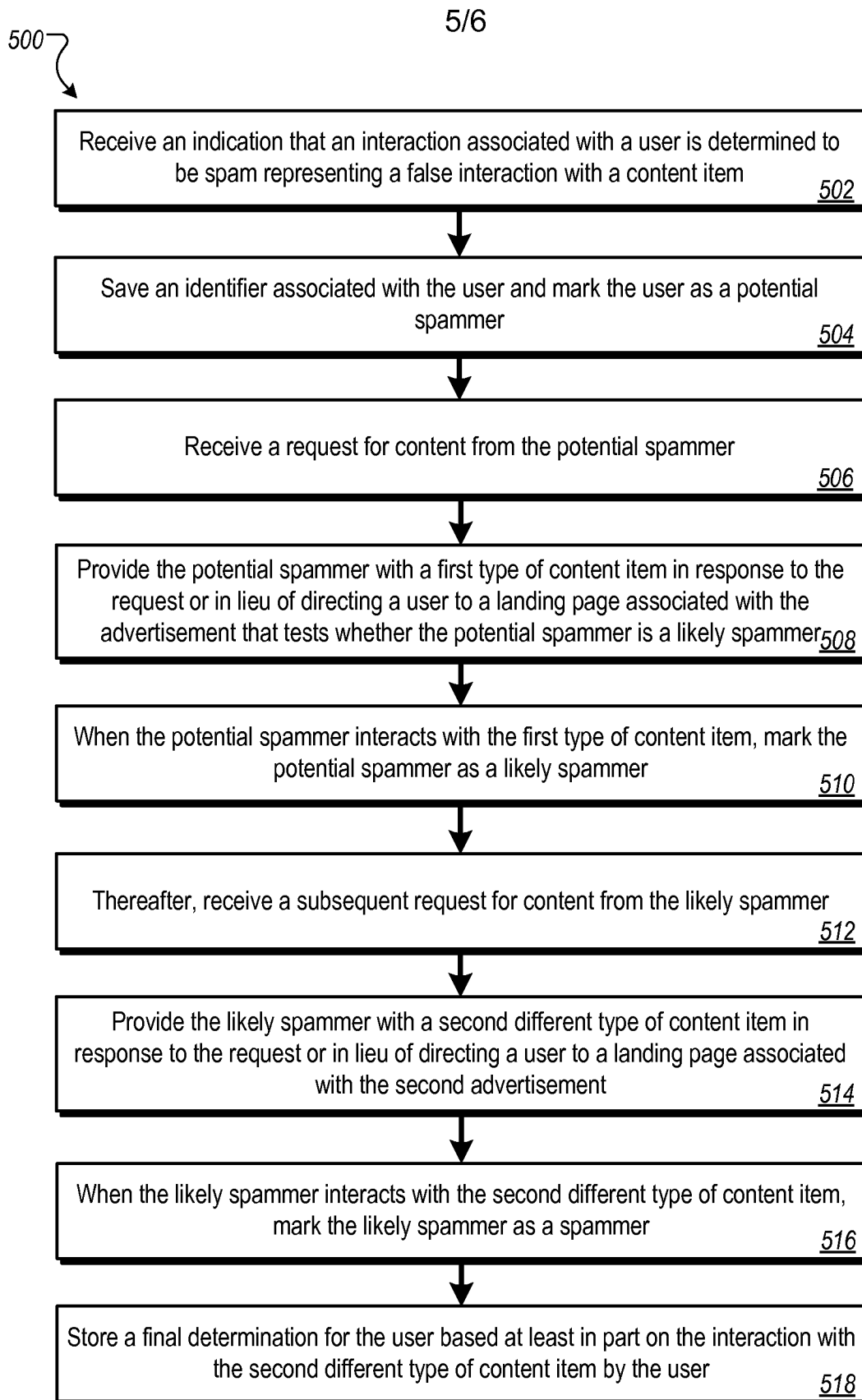


FIG. 5

6/6

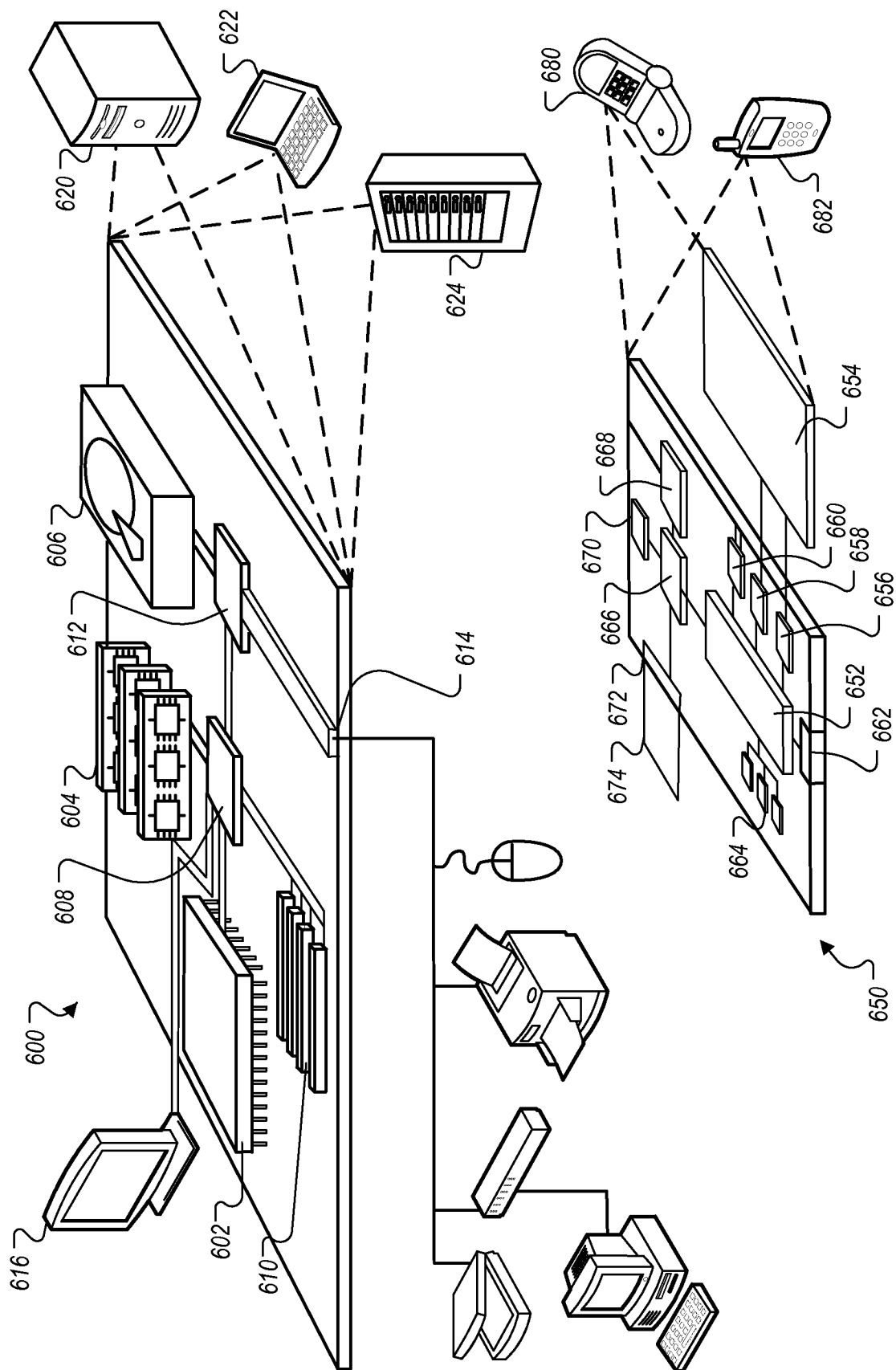


FIG. 6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2013/039738**A. CLASSIFICATION OF SUBJECT MATTER****G06F 15/16(2006.01)i**

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)

G06F 15/16; G06Q 50/30

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

Korean utility models and applications for utility models

Japanese utility models and applications for utility models

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

eKOMPASS(KIPO internal) & Keywords:spam detector , spammer, potential spammer, false positive rati, advertisement

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	KR 10-1201045 B1 (MICROSOFT CORP.) 14 November 2012 See abstract; paragraphs [24]-[29], [55]-[56], [63]; claims 1, 35, 60, 61 and figure 2.	1-20
A	US 2009-0216841 A1 (CHOI JAESIK et al.) 27 August 2009 See abstract; paragraphs [15]-[17], [33]-[35]; claims 1, 14 and figures 3, 4.	1-20
A	Amit A. Amleshwaram et al, "CATS: Characterizing Automation of Twitter Spammers", COMSNETS, 2013 Fifth International Conference, January 7-10, 2013 See Sections II.A and II.B.	1-20
A	Fabrizio Benevenuto et al, "Detecting Spammers on Twitter", Seventh annual Collaboration, Electronic messaging, AntiAbuse and Spam Conference, July 13-14, 2010 See Sections 4 and 5.	1-20

☐ 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 application or patent 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 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

26 November 2013 (26.11.2013)

Date of mailing of the international search report

27 November 2013 (27.11.2013)

Name and mailing address of the ISA/KR

Korean Intellectual Property Office
189 Cheongsu-ro, Seo-gu, Daejeon Metropolitan City,
302-701, Republic of Korea

Facsimile No. +82-42-472-7140

Authorized officer

HONG, Kyoung Ah

Telephone No. +82-42-481-5668



INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/US2013/039738

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
KR 10-1201045 B1	14/11/2012	CN 1573783 A EP 1496655 A2 EP 1496655 A3 JP 04-694146B2 JP 2005-011325A US 2005-0021649 A1 US 7711779 B2	02/02/2005 12/01/2005 05/10/2005 08/06/2011 13/01/2005 27/01/2005 04/05/2010
US 2009-0216841 A1	27/08/2009	US 7849146 B2	07/12/2010