(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
22 May 2014 (22.05.2014)

WIPO | PCT

(10) International Publication Number
WO 2014/075571 A1

(51) International Patent Classification:
G06F 9/45 (2006.01)

(21) International Application Number:
PCT/CN2013/086537

(22) International Filing Date:
5 November 2013 (05.11.2013)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
201210453701.1 13 November 2012 (13.11.2012) CN

(71) Applicant: **TENCENT TECHNOLOGY (SHENZHEN) COMPANY LIMITED** [CN/CN]; Room 403, East Block 2, SEG Park, Zhenxing Road, Futian District, Shenzhen, Guangdong 518000 (CN).

(72) Inventor: **TAO, Sinan**; Room 403, East Block 2, SEG Park, Zhenxing Road, Futian District, Shenzhen, Guangdong 518000 (CN).

(74) Agent: **DEQI INTELLECTUAL PROPERTY LAW CORPORATION**; 7/F, Xueyuan International Tower, NO.1 Zhichun Road, Haidian District, Beijing 100083 (CN).

(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, IR, IS, JP, KE, KG, 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, SA, 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, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

(54) **Title**: METHOD AND DEVICE FOR DETECTING MALICIOUS URL

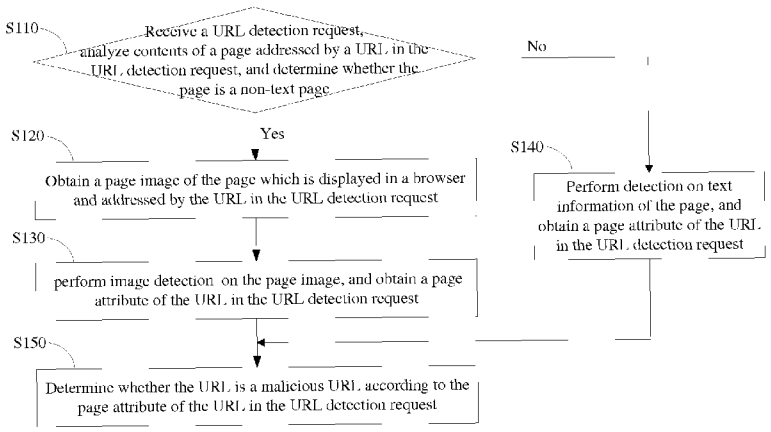

Figure 1

(57) **Abstract**: Examples of the present disclosure provide a method and device for detecting a malicious URL, the method includes: a URL detection request is received, contents of a page addressed by a URL in the URL detection request are analyzed, and it is de - termined that whether the page is a non-text page; when the page is a non-text page,a page image of the page, which is displayed in a browser and addressed by the URL in the URL detection request, is obtained,image detection is performed on the page image, and a page attribute of the URL in the URL detection request is obtained,whether the URL is a malicious URL is determined based on the page attribute of the URL in the URL detection request. Technical solutions based on the example of the present disclosure may ef - fectively identify not only a malicious URL of which the whole webpage is an image, but also a malicious URL evading detection through various encryption methods, malicious interference, and so on. Thus the security of users when obtaining online information may be further ensured.

WO 2014/075571 A1

METHOD AND DEVICE FOR DETECTING MALICIOUS URL

PRIORITY STATEMENT

[0001] This application claims the benefit of Chinese Patent Application No. 201210453701.1, filed on November13, 2012, the disclosure of which is incorporated herein in its entirety by reference.

FIELD

[0002] The present disclosure relates to (Uniform/ Universal Resource Locator, URL) detection field, and more particularly, to a method and device for detecting a malicious URL.

BACKGROUND

[0003] The network enriches people's life, but more and more pornography, fraud, phishing web sites also emerge, and bring a serious threat to the security of the majority of Internet users when obtaining information on the Internet. Thus a detection engine for identifying malicious URLs is needed.

[0004] An existing URL cloud detection engine may effectively identify and prompt whether a URL accessed by a user has a malicious behavior. After the user inputs a URL to be accessed and before the browser displays corresponding page content, it is necessary for the URL cloud detection engine to obtain malicious attributes of the URL to be accessed by the user from a cloud detection center, identify whether the URL to be accessed by the user has a malicious behavior, and provide relevant prompts based on the identification result. Due to the malicious web sites are variant, the URL cloud detection engine must possess fast, efficient and accurate characteristics, so as to ensure that the malicious web sites may be timely and accurately found.

[0005] The identification for malicious attributes by existing URL cloud detection engine may be performed through text information of page DOM and BOM object, and using machine learning manner, such as Bayes classifier/ keyword filtering and similarity matching.Although above technology may effectively identify text-based malicious fraud web site, the technology may not effectively identify non-text web content.

[0006] Moreover, the malicious pages in the prior art may evade the killing of detection engine through the following means.

[0007] (1) Text content is converted into an image. The contents of the whole page are made into an image, thus the killing is fought against through the manner that whole page is an image.

[0008] (2) Plaintext is encrypted and hidden. Since current detection engine mainly relies on the text information of the page, malicious webpage editors process the text information of a plaintext using encryption technology. When encountering an encrypted string without any semantics, an identification module of the detection engine cannot effectively identify the malicious webpage.

[0009] (3) Streaming media is used to fight against the detection engine. In order to prevent from being identified by current detection technology, in the existing malicious webpage, text information is hidden and displayed in a streaming media, such as a Flash. Thus the killing of existing detection technology may be evaded effectively.

[0010] (4) Normal text information is adopted to interfere with the killing of a detection engine. In order to evade the killing of existing detection technology, a large amount of normal text which is not displayed may be added to page contents to interfere with the identification module.

[0011] Therefore, how to efficiently and accurately detect the malicious URL has become a difficult problem and challenge for detection technology.

SUMMARY

[0012] According to examples of the present disclosure, a method and device for detecting a maliciousURL is provided to efficiently and accurately detect a malicious URL, and protect the security of users when obtaining online information.

[0013] The method for detecting a maliciousURL provided by an example of the present disclosure includes:receiving a URL detection request, analyzing a page addressed by a URL in the URL detection request, and determining whether the page is a non-text page; when determining that the page is a non-text page,obtaining a page image of the page, which

isdisplayed in a browser and addressed by the URL in the URL detection request; performing image detection on the page image, and obtaining a page attribute of the URL in the URL detection request; determining whether the URL is a malicious URL according to the page attribute of the URL in the URL detection request.

[0014] The device for detecting a maliciousURL provided by an example of the present disclosure includes:a page analyzing module, configured toreceive a URL detection request, analyze a page addressed by a URL in the URL detection request, and determine whether the page is a non-text page; anda page attribute identifying module, configured to, when the page analyzing module determines thatthe page is a non-text page, obtain a page image of the page which is displayed in a browser and addressed by the URL in the URL detection request, perform image detection on the page image, obtain a page attribute of the URL in the URL detection request, and determine whether the URL is a malicious URL according to the page attribute of the URL in the URL detection request.

[0015] The device for detecting a malicious URL provided by another example of the present disclosure includes:a memory and aprocessor in communication with the memory;the memory store a group of instructions which may be executed by the processor, and the instructions comprise:a page analyzing instruction, toindicate receiving a URL detection request, analyzing contents of a page addressed by a URL in the URL detection request, and determining whether the page is a non-text page; anda page attribute identifying instruction, to indicate, when the page analyzing module determines thatthe page is a non-text page, obtaining a page image of the page which is displayed in a browser and addressed by the URL in the URL detection request, performing image detection on the page image, obtaining a page attribute of the URL in the URL detection request, and determining whether the URL is a malicious URL according to the page attribute of the URL in the URL detection request.

[0016] As can be seen from the above technical solutions of the present disclosure, contents of the page addressed by the URL in the URL detection request are analyzed. When it is determined that the page is a non-text page, a page snapshot is performed on the page which is addressed by the URL and displayed in the background of the browser, and snapshotted page image is detected to obtain a page attribute of the URL in the URL detection request. When it is determined that the page is a text page, text of the page is detected to obtain a page attribute of the URL in the URL detection request. Subsequently, whether the URL is a

malicious URL is determined according to the page attribute of the URLin the URL detection request. Method for detecting a malicious URL based on the example of the present disclosure may effectively identify not only a malicious URL of which the whole webpage is an image, but also a malicious URL evading detection through various encryption methods, malicious interference, and so on. Thus the security of users when obtaining online information may be further ensured.

BRIEF DESCRIPTION OF THE DRAWINGS

[0017] Features of the present disclosure are illustrated by way of example and not limited in the following figures, in which like numerals indicate like elements, in which:

[0018] Figure 1 is a flow diagram illustrating a method for detecting a maliciousURLbased on an example of the present disclosure.

[0019] Figure 2 is a flowdiagram illustrating a procedure for analyzing a page addressed by a URL in a URL detection request in the method for detecting a maliciousURL based on an example of the present disclosure.

[0020] Figure 3 is a flow diagram illustrating a procedureoffor processing a non-textpage addressed by a URL in the URL detection request in the method for detecting a maliciousURL based on an example of the present disclosure.

[0021] Figure 4 is a schematic diagram illustrating a device for detecting a maliciousURLbased on an example of the present disclosure.

[0022] Figure 5 is a schematic diagram illustrating a page analyzing module of the device for detecting a maliciousURLbased on an example of the present disclosure.

[0023] Figure 6 is a schematic diagram illustrating a page attribute identifying module of the device for detecting a maliciousURLbased on an example of the present disclosure.

[0024] Figure 7 is a schematic diagram illustrating animage detection unit of the page attribute identifying moduleshown in figure 6 based on an example of the present disclosure

[0025] Figure 8 is a schematic diagram illustrating a device for detecting a malicious URLbased on another example of the present disclosure.

DETAILED DESCRIPTION

[0026] Reference will now be made in detail to examples, which are illustrated in the accompanying drawings. In the following detailed description, numerous specific details are set forth in order to provide a thorough understanding of the present disclosure. Also, the figures are illustrations of an example, in which modules or procedures shown in the figures are not necessarily essential for implementing the present disclosure. In other instances, well-known methods, procedures, components, and circuits have not been described in detail so as not to unnecessarily obscure aspects of the examples.As used herein, the term “includes” means includes but not limited to, the term “including” means including but not limited to. The term “based on” means based at least in part on. In addition, the terms “a” and “an” are intended to denote at least one of a particular element.

[0027] Figure 1 is a flow diagram illustrating a method for detecting a maliciousURLbased on an example of the present disclosure. As shown in figure 1, the methodincludes the following processes.

[0028] In block S110, a URL detection request is received, a page addressed by a URL in the URL detection request is analyzed, and it is determined that whether the page is a non-text page; when the page is a non-text page, proceed with block S120; otherwise, proceed with block S140.

[0029] The URL in the URL detection request may be a URL directly inputted by a user, or may be a URL generated after the user clicks a hyperlink. When the URL is received, it is possible to perform a preliminary analysis and filtration on the URL, and report suspicious URL, initiate a URL detection request. After the URL detection request is received, a page addressed by the URL in the URL detection request may be analyzed, so as to determine that whether the page is a text page or non-text page.

[0030] In block S120, a page image of the page, which isdisplayed in a browser and addressed by the URL in the URL detection request, is obtained.

[0031] After it is determined that the page addressed by the URL in the URL detection request is a non-text page, the browser is controlled to display the page in the background, and obtain a snapshot of the displayed page, thus a page image of the page which is displayed in a browser and addressed by the URL in the URL detection request is obtained.The snapshot may be achieved through an open source WebKit browser kernel, of course, also may be achieved using other methods, for example, through the Firefox, the IE or other browser kernel.

[0032] In block S130, image detection is performed on the page image, and a page attribute of the URL in the URL detection request is obtained, then proceed with block S150.

[0033] In block S140, detection is performed on text information of the page, and a page attribute of the URL in the URL detection request is obtained, then proceed with block S150.

[0034] In block S150, whether the URL is a malicious URL is determined according to the page attribute of the URL in the URL detection request.

[0035] Whether through text content encryption, image conversion, streaming media or other information hiding technology, the page of a malicious URL will still be displayed in the browser, and conducts effective phishing scams. Therefore, in the examples of the present disclosure, contents of the page addressed by the URL in the URL detection request are analyzed.When it is determined that the page is a non-text page, a page snapshot is performed on the page which is addressed by the URL and displayed in the background of the browser, and snapshotted page image is detected to obtain a page attribute of the URL in the URL detection request. When it is determined that the page is a text page, text of the page is detected to obtain a page attribute of the URL in the URL detection request. Subsequently, whether the URL is a malicious URL is determined according to the page attribute of the URLin the URL detection request. Method for detecting a malicious URL based on the example of the present disclosure may effectively identify not only a malicious URL of which the whole webpage is an image, but also a malicious URLevading detection through various encryption methods, malicious interference, and so on.Thus the security of users when obtaining online information may be further ensured.

[0036] In general, other web pages may be nested in a web page browsed by a user. Therefore, in above mentioned block S110, when the URL detection request is received, URL crawler grabbing may be performed starting from the page addressed by a URL in the URL detection request, and a HTML document corresponding to each URL may be generated. For example, the URL in the URL detection request may be taken as an initial URL, and crawler grabbing is performed on contents of the page addressed by the initial URL through a web crawler, new URL addresses are continuously extracted from the page and putted in a queue until a terminating condition, such as crawler grabbing depth and breadth, preconfigured by the web crawler is satisfied. All URLs grabbed by the web crawler areanalyzed and filtered, and then indexes are established for URLs after beinganalyzed and filtered. The URLs with established indexes may be stored, such that subsequent query and retrieval may be facilitated.

[0037] Subsequently, page analysis may be performed on each HTML document grabbed by the network crawler, and a Document Object Model (DOM) tree corresponding to each URL may be generated; data for determining a page attribute is obtained from each DOM tree, and it is determined whether the content feature of a page addressed by each URL is non-text content according to obtained data. Finally, it is determined whether the page addressed by the URL in the URL detection request is a non-text page according to content features of all URLs.The data for determining a page attribute may include executable JavaScript (JS), a page title and copyright information.

[0038] For example, as shown in figure 2, above mentioned block S110 may include the following processes.

[0039] In block S111, a URL detection request is received, crawler grabbing is performed ona page addressed by the URL (referred to as the initial URL) in the URL detection request, and a HTML document is generated.

[0040] Specifically, a network crawler may perform the crawler grabbing according to a preset grabbing breadth.

[0041] In block S112, page analysis is performed on the HTML document, and a DOM tree corresponding to the initial URL is generated; data for determining a page attribute is

obtained from the DOM tree, and it is determined whether the content feature of the page addressed by the initial URL is non-text content according to obtained data.

[0042] The data for determining a page attribute may include executable JavaScript (JS), a page title and copyright information.

[0043] Specifically, whether the content feature of the page addressed by the initial URL is non-text content may be determined according to label information of the obtained data. For example, when there is only little text information of non-HTML label or only HTML label information in obtained data, it is determined that content feature of the page addressed by the initial URL is non-text content.

[0044] In block S113, it is determined whether a terminating condition, such as presetcrawler grabbing depth, is satisfied, when the terminating condition is satisfied, proceed with block S119; otherwise, proceed with block S114.

[0045] In block S114, it is determined whether the DOM treecorresponding to the initial URL includes at least one sub-URL, when the DOM treecorresponding to the initial URL includes at least one sub-URL, proceed with block S115; otherwise, proceed with block S119.

[0046] In block S115, crawler grabbing is performed on a page addressed by each sub-URL in the DOM tree, and a HTML document corresponding to each sub-URL is generated.

[0047] Similarly, a network crawler may perform the crawler grabbing according to a preset grabbing breadth.

[0048] In block S116, page analysis is performed on each HTML documentcorresponding to a sub-URL, and a DOM tree corresponding to each URL is generated; data for determining a page attribute is obtained from each DOM tree corresponding to a sub-URL, and it is respectively determined whether the content feature of a page addressed by eachsub-URL is non-text content according to obtained data.

[0049] The data for determining a page attribute may include executable JavaScript (JS), a page title and copyright information.

[0050] Specifically, whether the content feature of the page addressed by current sub-URL is non-text content may be determined according to label information of obtained data corresponding to current sub-URL. For example, when there is only little text information of non-HTML label or only HTML label information in the obtained data, it is determined that content feature of the page addressed by current sub-URL is non-text content.

[0051] In block S117, it is determined whether a terminating condition, such as preset crawler grabbing depth, is satisfied, when the terminating condition is satisfied, proceed with block S119; otherwise, proceed with block S118.

[0052] In block S118, a DOM tree is extracted from DOM trees corresponding to sub-URLs in turn, and it is determined whether the DOM treecorresponding to a sub-URL includes at least one sub-URL, when the DOM treecorresponding to the sub-URL includes at least one sub-URL, return to block S115; otherwise, proceed with block S119.

[0053] In block S119, it is determined whether the page addressed by the initial URL in the URL detection request is a non-text page according to content features of pages addressed by all sub-URLs and the content feature of the page addressed by the initial URL.

[0054] Specifically, when content features of pages addressed by all sub-URLs and the content feature of the page addressed by the initial URL are all non-text content, it may be determined that the page addressed by the initial URL is a non-text page; otherwise, it may be determined that the page addressed by the initial URL is a text page.

[0055] Furthermore, as shown in figure 3, above mentioned block S130 may include the following processes.

[0056] In block S131, the page image is matched with a pre-stored seed page image.

[0057] The seed page image is a pre-stored page image addressed by a malicious URL. The block S131 may include the followings.

[0058] Firstly, an image feature of a page image to be matched (hereinafter referred to as the "target image") are extracted.

[0059] For example, the size of the target image may be reduced to 64 pixels. The details of the target image may be removed, only the basic information, such as structure and chiaroscuro, may beretained, and the image difference result from different scales may be abandoned.

[0060] Subsequently, the extracted image feature is encoded.

[0061] For example, the reduced image may be converted to a 64 grayscale image.A gray average of the 64 pixels is calculated,the gray of each pixel is compared with the gray average, when the gray is greater than or equal to the gray average, the pixel is denoted as 1; when the gray is less than or equal to the gray average, the pixel is denoted as 0. Finally, a 64-bit integer is formed.

[0062] Finally, the encoded image feature is matched with pre-stored seed page images in a page image database, and a similarity, namely the matching degree, of the target imageis obtained.

[0063] The 64-bit integer is matched with pre-stored page images in the page image database, when the number of different bits does not exceed a first threshold, it indicates that the two images are very similar; and when the number of different bits exceeds a second threshold, it indicates that the two images are different. In the example of the present disclosure, the first threshold is 5, and the second threshold is 10.

[0064] In block S132, it is determined that whether the matching degree of the page image and the pre-stored seed page image is greater than or equal to a preset value, when the matching degree of the page image and the pre-stored seed page image is greater than or equal to the preset value, proceed with block S133; otherwise, proceed with block S134.

[0065] The matching degree refers to the similarity degree when the target image is matched with the seed page image, namely the number of different bits of the two images. The smaller the number of different bits of the two images, the higher the matching degree of the two images, which represents that the URL of target image is a malicious URL.In the present example, the preset value is a matching degree corresponding to the number 5of different bits.

[0066] In block S133, it is determined that a page attribute of the URL is a malicious attribute, proceed with block S150.

[0067] In block S134, image-text identification is performed on the page image, and text informationof the page is obtained, proceed with block S140.

[0068] Furthermore, the method for performing detection on text information of the page in above mentioned block 140 may include one or more of the text segmentation, text similarity matching and machine identification.

[0069] The text segmentation refers to that segmentation is performed on text content of the page, and semantic information of the text information of the page is obtained. The text similarity matching refers to that similarity matching is performed on the semantic information obtained after the text segmentation and pre-stored text information of a malicious page, and a matching result is outputted. The machine identification refers to that detection is performed on the semantic information obtained after the text segmentation through a machine learning method, such as the Bayes classifier, a keyword model and a decision tree, and then a detection result is outputted.

[0070] Figure 4 is a schematic diagram illustrating a device for detecting a malicious URLbased on an example of the present disclosure.As shown in figure 4, the device may include a page analyzing module 110 and a page attribute identifying module 120.

[0071] The page analyzing module 110 is configured to receive a URL detection request, analyze a page addressed by a URL in the URL detection request, and determine whether the page is a non-text page.

[0072] The page attribute identifying module 120 is configured to,when the page analyzing module 110 determines thatthe page is a non-text page, obtain a page image of the page, which is displayed in a browser and addressed by the URL in the URL detection request, perform image detection on the page image, obtain a page attribute of the URL in the URL detection request, and determine whether the URL is a malicious URL based on the page attribute of the URL in the URL detection request.

[0073] The URL in the URL detection request may be a URL directly inputted by a user, or may be a URL generated after the user clicks a hyperlink. When the URL is received, it is possible to perform a preliminary analysis and filtration on the URL, and report suspicious URL, initiate a URL detection request. After receiving the URL detection request, the page analyzing module 110 may analyze the page addressed by the URL in the URL detection request, so as to determine that whether the page is a text page or non-text page. When it is determined that the page is a non-text page, the page attribute identifying module 120 may control the browser to display the page in the background, and obtain a snapshot of the displayed page, thus a page image of the page which is displayed in a browser and addressed by the URL in the URL detection request is obtained.

[0074] Whether through text content encryption, image conversion, streaming media or other information hiding technology, the page of a malicious URL will still be displayed in the browser, and conducts effective phishing scams. Therefore, in the examples of the present disclosure, contents of the page addressed by the URL in the URL detection request are analyzed. When it is determined that the page is a non-text page, a page snapshot is performed on the page which is addressed by the URLanddisplayed in the background of the browser, and snapshotted page image is detected to obtain a page attribute of the URL in the URL detection request. When it is determined that the page is a text page, text of the page is detected to obtain a page attribute of the URL in the URL detection request. Subsequently, whether the URL is a malicious URL is determined according to the page attribute of the URLin the URL detection request. Method for detecting a malicious URL based on the example of the present disclosure may effectively identify not only a malicious URL of which the whole webpage is an image, but also a malicious URL evading detection through various encryption methods, malicious interference, and so on. Thus the security of users when obtaining online information may be further ensured.

[0075] Figure 5 is a schematic diagram illustrating a page analyzing module of the device for detecting a URLbased on an example of the present disclosure.As shown in figure 5, above mentioned page analyzing module 110 may include a crawler unit 111 and an analyzing unit 112.

[0076] The crawler unit 111 is configured to receive a URL detection request, perform crawler grabbingstarting from the page addressed by the URL in the URL detection request, and generate a HTML document corresponding to each URL.

[0077] The analyzing unit 112 is configured to perform page analysison each HTML document grabbed by the network crawler, and generate a DOM tree corresponding to each URL; obtain data for determining a page attribute from each DOM tree, and determine whether the content feature of a page addressed by each URL is non-text content according to obtained data, then determine whether the page addressed by the URL in the URL detection request is a non-text page according to content features of all URLs.For example, the analyzing unit may determine thatthe page addressed by the URL in the URL detection request is a non-text page when content features of pages addressed by all URLs are all non-text content; and determine that the page addressed by the URL in the URL detection request is a text page when not all content features of pages addressed by all URLs are non-text content.

[0078] In general, other web pages are nested in a web page browsed by a user. When the URL detection request is received, the URL in the URL detection request is taken as an initial URL. The crawler unit 111 may perform crawler grabbing on contents of the page addressed by the initial URL, and continuously extract new URL addresses from the page and put extracted URL addresses in the URL queue until a stopping condition, such as crawler grabbing depth and breadth, preconfigured by a web crawler is satisfied. All URLs grabbed by the crawler unit 111 areanalyzed and filtered, and then indexes are established for URLs after beinganalyzed and filtered. The URLs with established indexes may be stored, such that subsequent query and retrieval may be facilitated. The analyzing unit 112 may perform page analysis on each HTML document grabbed by the crawler unit 111, obtain data, such as executable JavaScript (JS), a page title and copyright information, for determining page attributes, and determine whether the content feature of a page addressed by each URL is non-text content according to obtained data, and then determine whether the page addressed by the URL in the URL detection request is a non-text page according to content features of all URLs.

[0079] Figure 6 is a schematic diagram illustrating a page attribute identifying module of the device for detecting a malicious URLbased on an example of the present disclosure.As shown

in figure 6, above mentioned page attribute identifying module 120 may include an image detection unit 121, an image identification unit 122, a text detection unit 123 and an attribute determination unit 124.

[0080] The image detection unit 121 is configured to match the page image with a pre-stored seed page image, when a matching degree of the page image and the pre-stored seed page image is greater than or equal to a preset value, determine a page attribute of the URL is a malicious attribute.

[0081] The image identification unit 122 is configured to, when the matching degree of the page image and the pre-stored seed page image is less than the preset value, perform image-text identification on the page image, and obtain text informationof the page.

[0082] The text detection unit 123 is configured to perform detection on the text information of the page, and obtain a page attribute of the URL.

[0083] The attribute determination unit 124 is configured to determine whether the URL is a malicious URL according to the page attribute of the URL in the URL detection request.

[0084] Furthermore, the text detection unit 123 is further configured to, when it is determined that the page addressed by the URL in the URL detection request is a text page, perform detection on text information of the page, and obtain a page attribute of the URL.

[0085] Furthermore, the method for detecting text information of the page by the text detection unit 123 may include one or more of the text segmentation, text similarity matching and machine identification.

[0086] Figure 7 is a schematic diagram illustrating theimage detection unit of the page attribute identifying moduleshown in figure 6 based on an example of the present disclosure.As shown in figure 7, above mentionedimage detection unit may further include a feature extracting subunit 1211, an encoding subunit 1212 and a matching subunit 1213.

[0087] The feature extracting subunit 1211 is configured to extract an image feature of the page image.

[0088] The encoding subunit 1212 is configured to encode extracted image feature.

[0089] The matching subunit 1213 is configured to match encoded image featurewith pre-stored seed page images in a page image database, and obtain a matching degree corresponding to the page image.

[0090] The above examples may be implemented by hardware, software, firmware, or a combination thereof. For example the various methods, processes and functional modules described herein may be implemented by a processor (the term processor is to be interpreted broadly to include a CPU, processing unit/module, ASIC, logic module, or programmable gate array, etc.). The processes, methods and functional modules may all be performed by a single processor or split between several processors; reference in this disclosure or the claims to a 'processor' should thus be interpreted to mean 'one or more processors'. The processes, methods and functional modules are implemented as machine readable instructions executable by one or more processors, hardware logic circuitry of the one or more processors or a combination thereof. The modules, if mentioned in the aforesaid examples, may be combined into one module or further divided into a plurality of sub-modules. Further, the examples disclosed herein may be implemented in the form of a software product. The computer software product is stored in a non-transitory storage medium and comprises a plurality of instructions for making an electronic device implement the method recited in the examples of the present disclosure.

[0091] For example, figure 8is a schematic diagram illustrating anotherstructure of the device for detecting a malicious URLbased on an example of the present disclosure. As shown in figure 8, the device may include a memory 810and a processor 820 in communication with the memory 810.

[0092] The memory 810may store a group of instructions which may be executed by the processor 820. The instructions may include a page analyzing instruction811 and a page attribute identifying instruction 812, which may be respectively executed by the processor 820 to respectively implement the operations of the page analyzing module 110 and the page attribute identifying module 120 mentioned above.

[0093] Furthermore, the page analyzing instruction 811 may further include a crawler instruction and an analyzing instruction, which may be respectively executed by the

processor 820 to respectively achieve the operations of the crawler unit 111 and the analyzing unit 112 mentioned above.

[0094] Similarly, the page attribute identifying instruction 812 may further include an image detection instruction, an image identification instruction, a text detection instruction and an attribute determination instruction, which may be respectively executed by the processor 720 to respectively achieve the operations of the image detection unit 121, an image identification unit 122, a text detection unit 123 and an attribute determination unit 124 mentioned above.

[0095] The image detection instruction 121 may further include a feature extracting instruction, an encoding instruction and a matching instruction, which may be respectively executed by the processor 720 to respectively implement the operations of thefeature extracting subunit 1211, the encoding subunit 1212 and the matching subunit 1213 mentioned above.

[0096] The foregoing description, for purpose of explanation, has been described with reference to specific examples. However, the illustrative discussions above are not intended to be exhaustive or to limit the present disclosure to the precise forms disclosed. Many modifications and variations are possible in view of the above teachings. The examples were chosen and described in order to best explain the principles of the present disclosure and its practical applications, to thereby enable others skilled in the art to best utilize the present disclosure and various examples with various modifications as are suited to the particular use contemplated.

CLAIMS

WHAT IS CLAIMED IS:

1. A method for detecting a malicious Uniform/ Universal Resource Locator, URL, comprising:

receiving a URL detection request, analyzing a page addressed by a URL in the URL detection request, and determining whether the page is a non-text page;

when determining that the page is a non-text page,obtaining a page image of the page, which isdisplayed in a browser and addressed by the URL in the URL detection request;

performingimage detection on the page image, and obtaining a page attribute of the URL in the URL detection request; and

determining whether the URL is a malicious URL according to the page attribute of the URL in the URL detection request.

2. The method according to claim 1, wherein theanalyzing a page addressed by a URL in the URL detection request, and determining whether the page is a non-text page, comprises:

performingcrawler grabbingstarting from the page addressed by the URL in the URL detection request, and generating a HTML document corresponding to each URL;

performing page analysis on each HTML document, and generating a Document Object Model, DOM, tree document corresponding to each URL;

obtaining data for determining a page attribute from each DOM tree, and determining whether content feature of a page addressed by each URL is non-text content according to the data; and

determiningwhether the page addressed by the URL in the URL detection request is a non-text page according to content features of pages addressed by all URLs.

3.The method according to claim 2, wherein determining whether the page addressed by the URL in the URL detection request is a non-text page according to content features of all URLs comprises:

whencontent features of pages addressed by all URLs are all non-text content,

determining that the page addressed by the URL in the URL detection request is a non-text page;

when not all content features of pages addressed by all URLs are non-text content,

determining that the page addressed by the URL in the URL detection request is a text page.

4.The method according to claim 2, whereinperforming image detection on the page image, and obtaining a page attribute of the URL in the URL detection request, comprises:

matching the page image with a pre-stored seed page image;

when a matching degree of the page image and the pre-stored seed page image is greater than or equal to a preset value, determining a page attribute of the URL is a malicious attribute;

when the matching degree of the page image and the pre-stored seed page image is less than the preset value, performing image-text identification on the page image, and obtaining text informationof the page; and

detecting the text information of the page, and obtaining a page attribute of the URL.

5.The method according to claim 4, wherein matching the page image with the pre-stored seed page image comprises:

extractingan image feature of the page image;

encoding the image feature;

matching encoded image featurewith pre-stored seed page images in a page image database, and obtaining a matching degreecorresponding to the page image.

6.The method according to claim 1, further comprising:

when determining that the page is a text page,detecting text information of the page, and obtaining a page attribute of the URL.

7.The method according to any one of claims4 to6, wherein the detecting text information of the page comprises:

detecting text information of the page adopting one or more of text segmentation, text similarity matching and machine identification.

8. A device for detecting a malicious Uniform/ Universal Resource Locator, URL, comprising:

a page analyzing module, configured toreceive a URL detection request, analyze contents of a page addressed by a URL in the URL detection request, and determine whether the page is a non-text page; and

a page attribute identifying module, configured to, when the page analyzing module determines thatthe page is a non-text page, obtain a page image of the page which is displayed in a browser and addressed by the URL in the URL detection request, perform image detection on the page image, obtain a page attribute of the URL in the URL detection request, and determine whether the URL is a malicious URL according to the page attribute of the URL in the URL detection request.

9. Thedeviceaccording to claim8, wherein the page analyzing module comprises:

a crawler unit, configured to receive a URL detection request, perform crawler grabbingstarting from the page addressed by the URL in the URL detection request, and generate a HTML document corresponding to each URL; and

ananalyzing unit, configured to perform page analysison each HTML document grabbed by the network crawler, and generate a DOM tree corresponding to each URL; obtain data for determining a page attribute from each DOM tree, and determine whether the content feature of a page addressed by each URL is non-text content according to obtained data, then determine whether the page addressed by the URL in the URL detection request is a non-text page according to content features of pages addressed by all URLs.

10. Thedeviceaccording to claim 9, wherein the analyzing unit determines thatthe page addressed by the URL in the URL detection request is a non-text page when content features of pages addressed by all URLs are all non-text content; and determines that the page addressed by the URL in the URL detection request is a text page when not all content features of pages addressed by all URLs are non-text content.

11. Thedeviceaccording to claim9, wherein the page attribute identifying module comprises:

an image detection unit, configured to match the page image with a pre-stored seed page image, when a matching degree of the page image and the pre-stored seed page image is greater than or equal to a preset value, determine a page attribute of the URL is a malicious attribute;

an image identification unit, configured to, when the matching degree of the page image and the pre-stored seed page image is less than the preset value, perform image-text identification on the page image, and obtain text informationof the page;

a text detection unit, configured to perform detection on the text information of the page, and obtain a page attribute of the URL; and

anattribute determination unit, configured to determine whether the URL is a malicious URL according to the page attribute of the URL in the URL detection request.

12. Thedeviceaccording to claim 11, wherein the image detection unitcomprises:

a feature extracting subunit, configured to extract an image feature of the page image;

an encoding subunit, configured to encode extracted image feature; and

amatching subunit, configured to match encoded image featurewith pre-stored seed page images in a page image database, and obtain a matching degree corresponding to the page image.

13. Thedeviceaccording to claim 11, wherein the text detection unit is further configured to, when it is determined that the page addressed by the URL in the URL detection request is a text page, perform detection on text information of the page, and obtain a page attribute of the URL.

14. Thedeviceaccording to any one of claims 11 to 13, wherein the text detection unit is configured toperform detection on text information of the page adopting one or more oftext segmentation, text similarity matching and machine identification.

15. A device for detecting a malicious Uniform/ Universal Resource Locator, URL, comprising: a memory and aprocessor in communication with the memory;

the memory store a group of instructions which may be executed by the processor, and the instructions comprise:

a page analyzing instruction, toindicate receiving a URL detection request, analyzing contents of a page addressed by a URL in the URL detection request, and determining whether the page is a non-text page; and

a page attribute identifying instruction, to indicate, when the page analyzing module determines thatthe page is a non-text page, obtaining a page image of the page which is displayed in a browser and addressed by the URL in the URL detection request, performing image detection on the page image, obtaining a page attribute of the URL in the URL detection request, and determining whether the URL is a malicious URL according to the page attribute of the URL in the URL detection request.

16. Thedeviceaccording to claim 15, wherein the page analyzing instruction comprises:

a crawler instruction, to indicate receiving a URL detection request, performing crawler grabbingstarting from the page addressed by the URL in the URL detection request, and generating a HTML document corresponding to each URL; and

an analyzing instruction, to indicate performing page analysison each HTML document grabbed by the network crawler, and generating a DOM tree corresponding to each URL; obtaining data for determining a page attribute

from each DOM tree, and determining whether the content feature of a page addressed by each URL is non-text content according to obtained data, then determining whether the page addressed by the URL in the URL detection request is a non-text page according to content features of pages addressed by all URLs.

17. Thedeviceaccording to claim16, wherein the page attribute identifying instructioncomprises:

an image detection instruction, to indicate matching the page image with a pre-stored seed page image, when a matching degree of the page image and the pre-stored seed page image is greater than or equal to a preset value, determining a page attribute of the URL is a malicious attribute;

an image identification instruction, to indicate, when the matching degree of the page image and the pre-stored seed page image is less than the preset value, performing image-text identification on the page image, and obtain text informationof the page;

a text detection instruction, to indicate performingdetection on the text information of the page, and obtaining a page attribute of the URL; and

anattribute determination instruction, to indicate determining whether the URL is a malicious URL according to the page attribute of the URL in the URL detection request.

18. Thedeviceaccording to claim 17, wherein the image detection instructioncomprises:

a feature extracting instruction, to indicate extracting an image feature of the page image;

an encoding instruction, to indicate encoding extracted image feature; and

a matching instruction, to indicate matchingencoded image featurewith pre-stored seed page images in a page image database, and obtaining a matching degree corresponding to the page image.

19. Thedeviceaccording to claim17, wherein the text detection instruction is further to indicate, when it is determined that the page addressed by the URL in the URL detection request is a text page, performingdetection on text information of the page, and obtaining a page attribute of the URL.

20. Thedeviceaccording to any one of claims17to19, wherein the text detection instruction is toindicate performingdetection on text information of the page adopting one or more oftext segmentation, text similarity matching and machine identification.

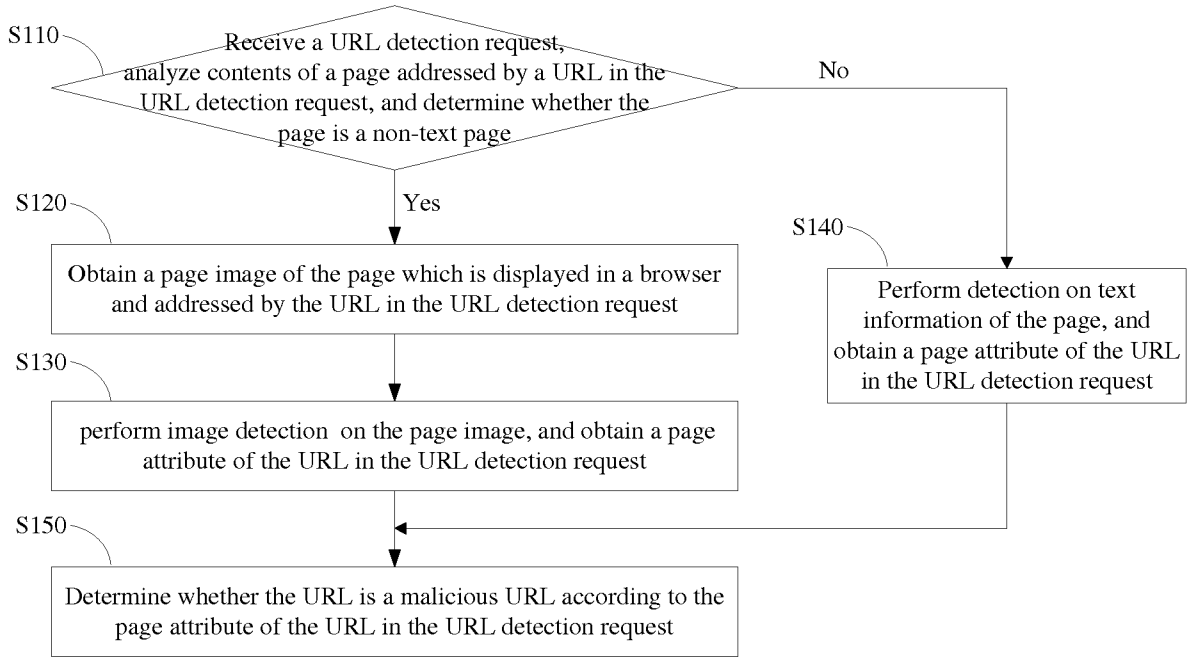
S110
Receive a URL detection request, analyze contents of a page addressed by a URL in the URL detection request, and determine whether the page is a non-text page
No
Yes
S120
Obtain a page image of the page which is displayed in a browser and addressed by the URL in the URL detection request
S140
Perform detection on text information of the page, and obtain a page attribute of the URL in the URL detection request
S130
perform image detection on the page image, and obtain a page attribute of the URL in the URL detection request
S150
Determine whether the URL is a malicious URL according to the page attribute of the URL in the URL detection request

Figure 1

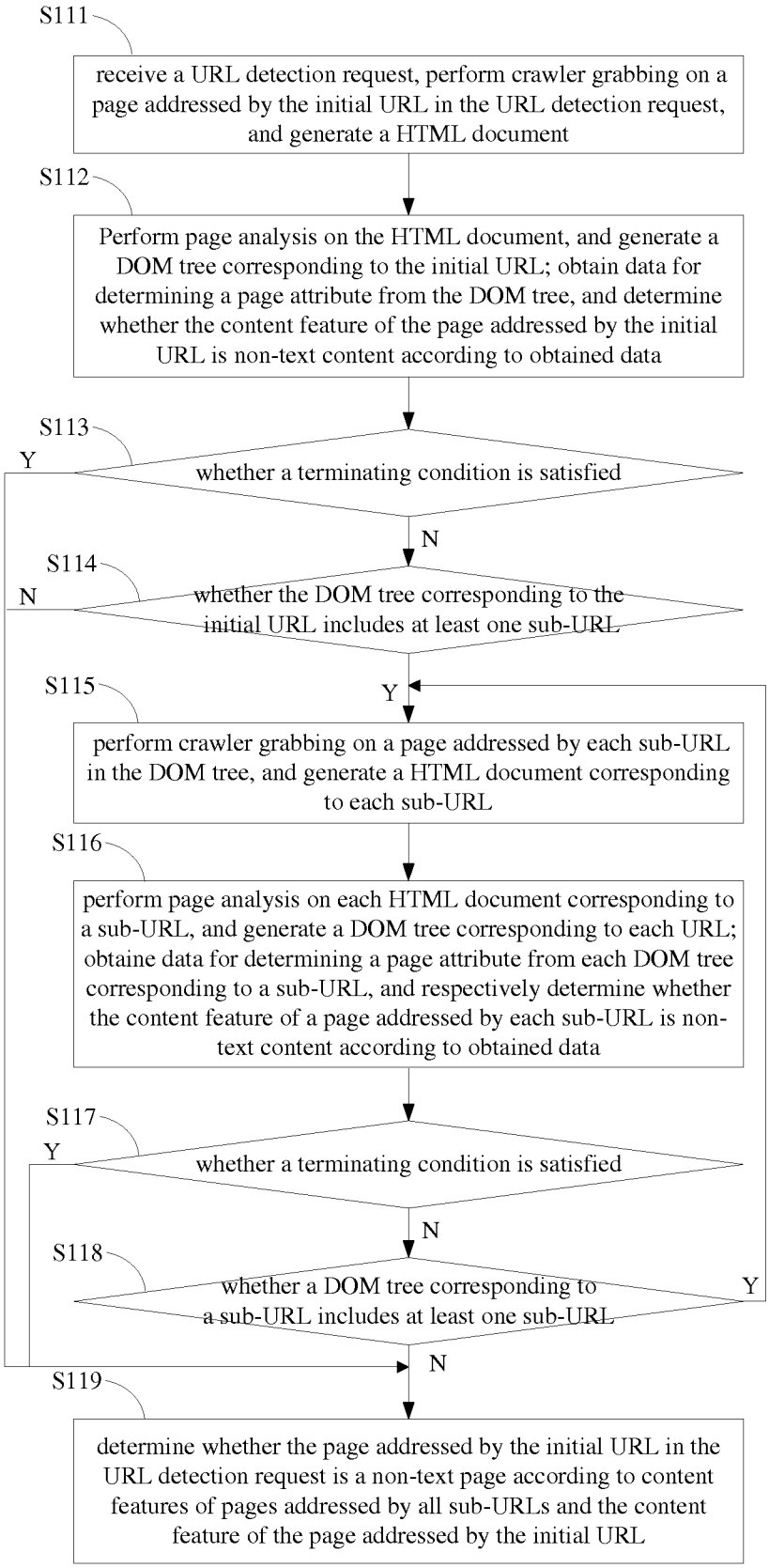
S111
receive a URL detection request, perform crawler grabbing on a page addressed by the initial URL in the URL detection request, and generate a HTML document
S112
Perform page analysis on the HTML document, and generate a DOM tree corresponding to the initial URL; obtain data for determining a page attribute from the DOM tree, and determine whether the content feature of the page addressed by the initial URL is non-text content according to obtained data
S113
whether a terminating condition is satisfied
Y
N
S114
whether the DOM tree corresponding to the initial URL includes at least one sub-URL
N
Y
S115
perform crawler grabbing on a page addressed by each sub-URL in the DOM tree, and generate a HTML document corresponding to each sub-URL
S116
perform page analysis on each HTML document corresponding to a sub-URL, and generate a DOM tree corresponding to each URL; obtaine data for determining a page attribute from each DOM tree corresponding to a sub-URL, and respectively determine whether the content feature of a page addressed by each sub-URL is non-text content according to obtained data
S117
whether a terminating condition is satisfied
Y
N
S118
whether a DOM tree corresponding to a sub-URL includes at least one sub-URL
Y
N
S119
determine whether the page addressed by the initial URL in the URL detection request is a non-text page according to content features of pages addressed by all sub-URLs and the content feature of the page addressed by the initial URL

Figure 2

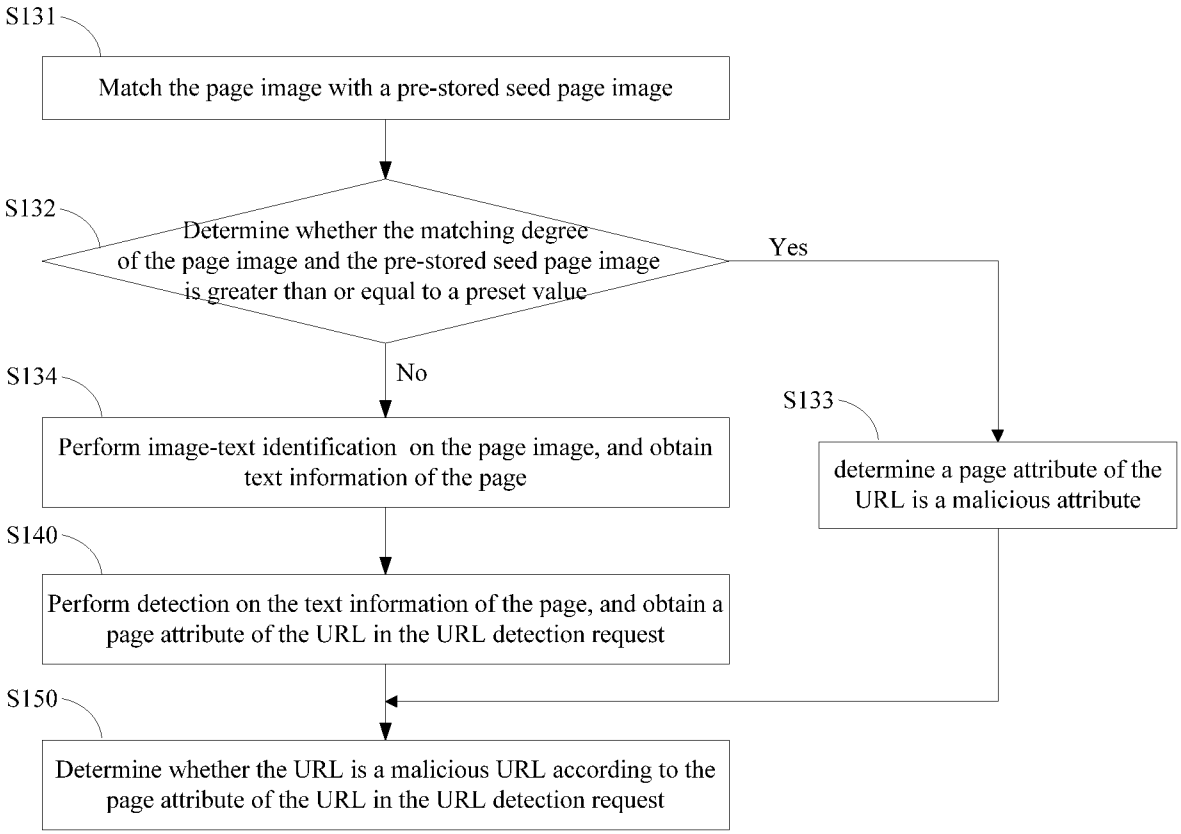

Figure 3

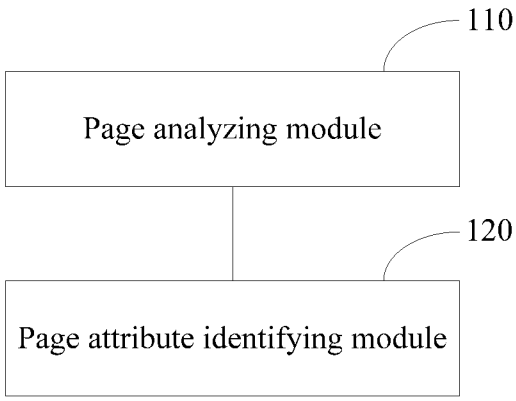

Figure 4

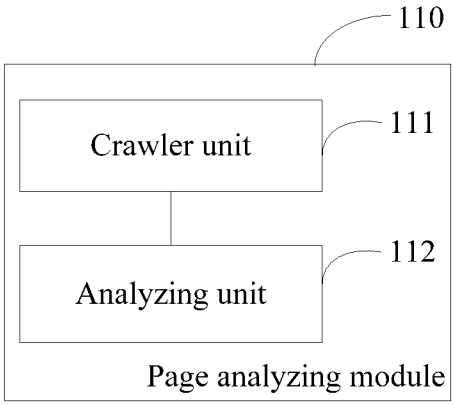

Figure 5

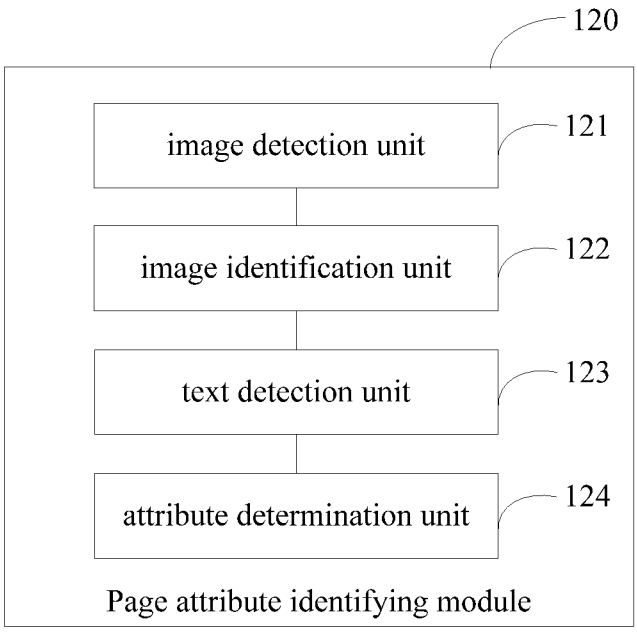

Figure 6

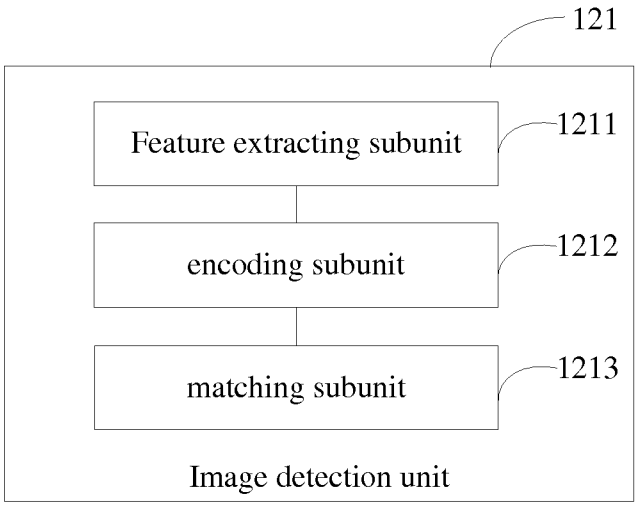

Figure 7

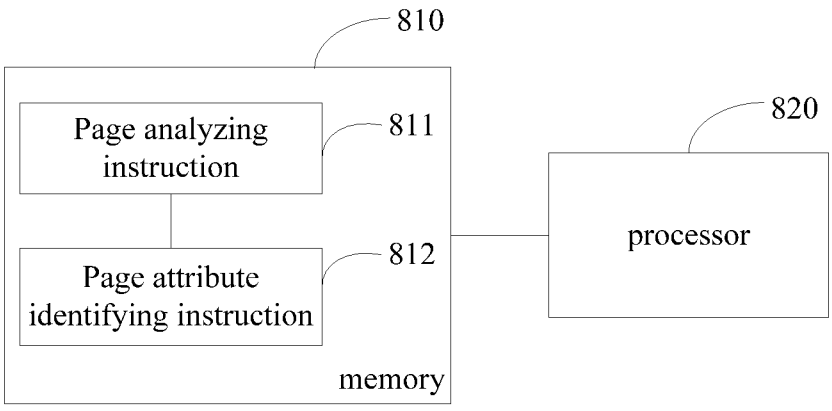

Figure 8

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2013/086537

A. CLASSIFICATION OF SUBJECT MATTER

G06F 9/45 (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)

IPC: G06F 9/-

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

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

CNABS,CNKI,VEN: malicious, web w page, image, detect, extract, fishing, network, internet, page w script, site, non w fishing, crawler, pornographic, cheat, URL, text, non w text, html, pure w text, plain w text, phishing

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 101145902 A (UNIV SOUTHEAST) 19 March 2008 (19.03.2008)	1-3,6-10,15-16
A	Abstract, description page 4 line 2 to line 18	4-5,11-14,17-20
Y	CN 102004779 A (BAIDU ON-LINE NETWORK TECHNOLOGY CO., LTD.)	1-3,6-10,15-16
A	06 April 2011 (06.04.2011) claim 17, claims 24 to 27	4-5,11-14,17-20
Y	CN 102063484 A (BEIJING ANTIY ELECTRONIC APPLIANCE CO., LTD.) 18 May 2011 (18.05.2011) abstract	2-3,9-10,16
Y	CN 101324888 A (BEIJING HENGJIN HENGTAI INFORMATION TECH) 17 December 2008 (17.12.2008) abstract	6-7

☐ 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 another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
16 January 2014 (16.01.2014)	**30 Jan. 2014 (30.01.2014)**
Name and mailing address of the ISA/CN The State Intellectual Property Office, the P.R.China 6 Xitucheng Rd., Jimen Bridge, Haidian District, Beijing, China 100088 Facsimile No. 86-10-62019451	Authorized officer Liu, Jia Telephone No. (86-10)62411655

Form PCT/ISA /210 (second sheet) (July 2009)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2013/086537

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 101145902 A	19.03.2008	CN 100583738 C	20.01.2010
CN 102004779 A	06.04.2011	CN 102004779 B	28.11.2012
CN 102063484 A	18.05.2011	CN 102063484 B	10.04.2013
CN 101324888 A	17.12.2008	None	