



- (51) **International Patent Classification:**
G06F 17/30 (2006.01)
- (21) **International Application Number:**
PCT/CN2013/085367
- (22) **International Filing Date:**
17 October 2013 (17.10.2013)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (30) **Priority Data:**
201210397725.X 18 October 2012 (18.10.2012) CN
- (71) **Applicant: TENCENT TECHNOLOGY (SHENZHEN) COMPANY LIMITED** [CN/CN]; Room 403, East Block 2, SEG Park, Zhenxing Road, Futian District, Shenzhen, Guangdong 518044 (CN).
- (72) **Inventors: LI, Yongjian;** Room 403, East Block 2, SEG Park, Zhenxing Road, Futian District, Shenzhen, Guangdong 518044 (CN). **ZHOU, Jiwen;** Room 403, East Block 2, SEG Park, Zhenxing Road, Futian District, Shenzhen, Guangdong 518044 (CN). **YU, Yang;** Room 403, East Block 2, SEG Park, Zhenxing Road, Futian District, Shenzhen, Guangdong 518044 (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 APPARATUS FOR DETECTING HIDDEN CONTENT OF WEB PAGE

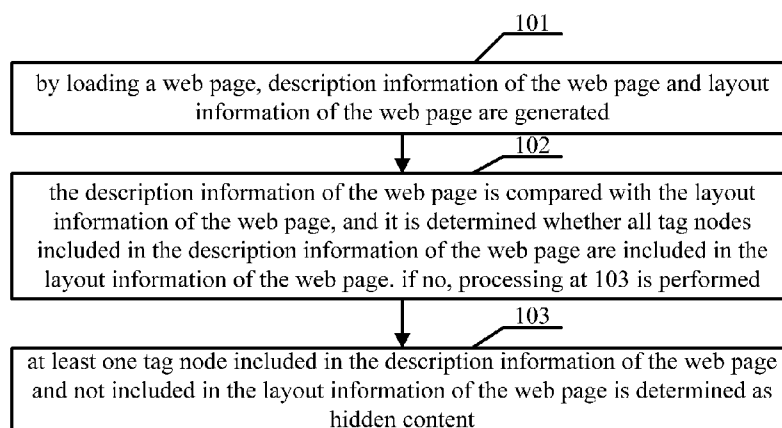


Fig. 1

(57) **Abstract:** According to an example, a web page is loaded and description information of the web page and layout information of the web page are generated, the description information of the web page is compared with the layout information of the web page, whether all tag nodes comprised in the description information of the web page are comprised in the layout information of the web page is determined, and it is determined at least one tag node, which being comprised in the description information of the web page and not contained in the layout information of the web page, as hidden content.



Method and Apparatus for Detecting Hidden Content of Web Page

Field of the Invention

The present invention relates to communication technologies, more particularly to, a
5 method and apparatus for detecting hidden content of a web page.

Background of the Invention

Currently, when hanging a hidden chain which is also called a black chain, a hacker
always configures an attribute of a Hypertext Markup Language (HTML) tag as display:
none, by using many modes. For example, the hacker may configure in the HTML,
10 configure in a Cascading Style Sheet (CSS) file via a CSS selector, or configure property
settings of a web page node via JavaScript which is a programming language.

When hanging the hidden chain, the hacker modifies page content of a normal
website. Hidden links pointing to other websites are added in a normal web page, and
thus the number of external links of a target website is improved. The external link is a
15 link pointing to a Uniform/Universal Resource Locator (URL) under a nonlocal domain
name. The hung hidden chain is hidden content of the web page. Traditional detection
modes are as follows.

In a static detection mode, the HTML, CSS and JavaScript commonly used by the
hacker are collected, general text string characteristics are extracted, and the general text
20 string characteristics are used to detect the hidden chain in the web page.

In a dynamic detection mode, JavaScript codes of the web page are executed by
using an open source JavaScript engine, if it is found that executed codes has a function
of hiding external links, it is determined that the page is hung with the hidden chain.

Because various hidden link codes may be generated by combining the HTML, CSS,
25 and JavaScript, and the JavaScript as a front programming language may encrypt
malicious code, so that it is easy to bypass the static detection. The hidden codes
encrypted by the JavaScript may be detected by using the dynamic detection, but no
processing is performed for the CSS and HTML, so that the web page with the hidden
content added by using a deformation HTML + CSS mode cannot be recognized.

Therefore, accuracy rate of detecting the hidden content by using each of the two detection modes is not high.

Summary of the Invention

The examples of the present invention provide a method and apparatus for detecting
5 hidden content of a web page, so as to improve accuracy rate of detecting the hidden content of the web page.

A method for detecting hidden content of a web page includes:
loading a web page;
generating description information of the web page and generating layout
10 information of the web page;
comparing the description information of the web page with the layout information of the web page according to a preconfigured rule;
determining whether all tag nodes comprised in the description information of the web page are contained in the layout information of the web page; and
15 determining at least one tag node, which being comprised in the description information of the web page and not being contained in the layout information of the web page, as hidden content.

An apparatus for detecting hidden content of a web page includes:
a loading unit, configured to load a web page and to generate description information
20 of a web page and generate layout information of the web page;
a comparing unit, configured to compare the description information of the web page with the layout information of the web page according to a preconfigured rule, and further to determine whether all tag nodes comprised in the description information of the web page are contained in the layout information of the web page; and
25 a content determining unit, configured to determine at least one tag node, which being comprised in the description information of the web page and not being contained in the layout information of the web page, as hidden content.

The purpose of configuring hidden content in the web page is not to display content to be hidden on the screen after the browser layouts the content. According to the
30 examples of the present application, the page is loaded by the layout engine, the

description information and layout information of the web page are generated, the description information are compared with the layout information, and thus detecting the hidden content is not affected by the types or encryption, thereby accurately detecting the hidden content and improving accuracy of detecting the hidden content of the web page.

5 Brief Description of the Drawings

Figure 1 is a schematic flowchart illustrating a method for detecting hidden content of a web page according to various examples of the present invention.

Figure 2 is a schematic flowchart illustrating a method for detecting hidden content of a web page according to various examples of the present invention.

10 Figure 3 is a schematic diagram illustrating a structure of an apparatus for detecting hidden content of a web page according to various examples of the present invention.

Figure 4 is a schematic diagram illustrating a structure of an apparatus for detecting hidden content of a web page according to various examples of the present invention.

15 Figure 5 is a schematic diagram illustrating a structure of an apparatus for detecting hidden content of a web page according to various examples of the present invention.

Detailed Description of the Invention

The examples of the present application provide the following technical solutions.

20 The following description is merely illustrative in nature and is in no way intended to limit the disclosure, its application, or uses. The broad teachings of the disclosure can be implemented in a variety of forms. Therefore, while this disclosure includes particular examples, the true scope of the disclosure should not be so limited since other modifications will become apparent upon a study of the drawings, the specification, and the following claims. For purposes of clarity, the same reference numbers will be used in the drawings to identify similar elements.

25 The terms used in this specification generally have their ordinary meanings in the art, within the context of the disclosure, and in the specific context where each term is used. Certain terms that are used to describe the disclosure are discussed below, or elsewhere in

the specification, to provide additional guidance to the practitioner regarding the description of the disclosure. The use of examples anywhere in this specification, including examples of any terms discussed herein, is illustrative only, and in no way limits the scope and meaning of the disclosure or of any exemplified term. Likewise,
5 the disclosure is not limited to various embodiments given in this specification.

Reference throughout this specification to "one embodiment," "an embodiment," "specific embodiment," or the like in the singular or plural means that one or more particular features, structures, or characteristics described in connection with an embodiment is included in at least one embodiment of the present disclosure. Thus, the
10 appearances of the phrases "in one embodiment" or "in an embodiment," "in a specific embodiment," or the like in the singular or plural in various places throughout this specification are not necessarily all referring to the same embodiment. Furthermore, the particular features, structures, or characteristics may be combined in any suitable manner in one or more embodiments.

15 As used in the description herein and throughout the claims that follow, the meaning of "a", "an", and "the" includes plural reference unless the context clearly dictates otherwise. Also, as used in the description herein and throughout the claims that follow, the meaning of "in" includes "in" and "on" unless the context clearly dictates otherwise.

As used herein, the terms "comprising," "including," "having," "containing,"
20 "involving," and the like are to be understood to be open-ended, i.e., to mean including but not limited to.

As used herein, the phrase "at least one of A, B, and C" should be construed to mean a logical (A or B or C), using a non-exclusive logical OR. It should be understood that one or more steps within a method may be executed in different order (or concurrently)
25 without altering the principles of the present disclosure.

As used herein, the term "module" may refer to, be part of, or include an Application Specific Integrated Circuit (ASIC); an electronic circuit; a combinational logic circuit; a field programmable gate array (FPGA); a processor (shared, dedicated, or group) that executes code; other suitable hardware components that provide the described
30 functionality; or a combination of some or all of the above, such as in a system-on-chip.

The term module may include memory (shared, dedicated, or group) that stores code executed by the processor.

The term “code”, as used herein, may include software, firmware, and/or microcode, and may refer to programs, routines, functions, classes, and/or objects. The term “shared”, as used herein, means that some or all code from multiple modules may be executed using a single (shared) processor. In addition, some or all code from multiple modules may be stored by a single (shared) memory. The term “group”, as used herein, means that some or all code from a single module may be executed using a group of processors. In addition, some or all code from a single module may be stored using a group of memories.

The systems and methods described herein may be implemented by one or more computer programs executed by one or more processors. The computer programs include processor-executable instructions that are stored on a non-transitory tangible computer readable medium. The computer programs may also include stored data. Non-limiting examples of the non-transitory tangible computer readable medium are nonvolatile memory, magnetic storage, and optical storage.

The description will be made as to the various embodiments in conjunction with the accompanying drawings in FIGS. 1-5. It should be understood that specific embodiments described herein are merely intended to explain the present disclosure, but not intended to limit the present disclosure. In accordance with the purposes of this disclosure, as embodied and broadly described herein, this disclosure, in one aspect, relates to method and apparatus for detecting hidden content of a web page.

Examples of mobile terminals that can be used in accordance with various embodiments include, but are not limited to, a tablet PC (including, but not limited to, Apple iPad and other touch-screen devices running Apple iOS, Microsoft Surface and other touch-screen devices running the Windows operating system, and tablet devices running the Android operating system), a mobile phone, a smartphone (including, but not limited to, an Apple iPhone, a Windows Phone and other smartphones running Windows Mobile or Pocket PC operating systems, and smartphones running the Android operating system, the Blackberry operating system, or the Symbian operating system), an e-reader

(including, but not limited to, Amazon Kindle and Barnes & Noble Nook), a laptop computer (including, but not limited to, computers running Apple Mac operating system, Windows operating system, Android operating system and/or Google Chrome operating system), or an on-vehicle device running any of the above-mentioned operating systems
5 or any other operating systems, all of which are well known to one skilled in the art.

Figure 1 is a schematic flowchart illustrating a method for detecting hidden content of a web page according to various examples of the present invention. As shown in Figure 1, the method includes the following processing.

At 101, by loading a web page, description information of the web page and layout
10 information of the web page are generated.

According to an example, the web page is loaded by using a layout engine, and before loading the web page by using the layout engine, a user-agent character string of the layout engine is configured as a search engine.

Currently, many hackers determine whether a web page is visited by a common user
15 or by a search engine crawler according to the user-agent. The hacker only returns a page with Search Engine Optimization (SEO) to the search engine crawler, and returns normal content to the common user. It is different for network managers to detect abnormal states of the web site. According to the example, when the user-agent of the layout engine is configured as the search engine, the above problem is solved, and accuracy of detecting
20 the hidden content of the web page is further improved.

When the web page is loaded by using the layout engine, a Uniform / Universal Resource Locator (URL) is loaded by using the layout engine, returned HTML codes are parsed, and other elements in a page corresponding to the returned HTML codes are loaded. The other elements may be CSS, JavaScript and etc., which is not limited in the
25 examples of the present invention.

According to an example, the layout engine is an open-source browser layout engine, e.g. a webkit, the description information of the web page is a Document Object Model (DOM) tree; the layout information of the web page is a render tree. It should be noted that, the layout engine is a tool for configuring the layout of the web page and is not

limited to the webkit according to various examples. Similarly, according to various examples, the description information and the layout information are not limited to the DOM tree and the render tree respectively.

Accordingly, at 101, the web page is loaded by using the webkit, and the DOM tree
5 and the render tree are generated.

At 102, the description information of the web page is compared with the layout information of the web page, and it is determined whether all tag nodes included in the description information of the web page are included in the layout information of the web page.

10 If not all tag nodes included in the description information of the web page are included in the layout information of the web page, processing at 103 is performed.

If all tag nodes included in the description information of the web page are included in the layout information of the web page, it indicates that the web page does not have hidden content, and the procedure is terminated.

15 According to an example, at 102, when the layout engine is the webkit, the description information of the web page is the DOM tree, and the layout information of the web page is the render tree, the DOM tree is compared with the render tree and it is determined whether all tag nodes included in the DOM tree are included in the render tree.

20 Further, the description information of the web page is compared with the layout information of the web page, after the description information of the web page and the layout information of the web page are generated and a preset delay time period expires.

According to an example, the preset delay time period may be configured according to time period needed for e.g. loading the web page, such as 5 second, which is not
25 limited in the examples of the present application.

At 103, at least one tag node included in the description information of the web page and not contained in the layout information of the web page is determined as hidden content.

According to an example, at 103, when the layout engine is the webkit, the description information of the web page is the DOM tree and the layout information of the web page is the render tree, if not all tag nodes of the DOM tree are included in the render tree, at least one tag node included in the DOM tree and not contained in the render tree is determined as the hidden content.

The purpose of configuring hidden content in the web page is not to display content to be hidden on the screen after the browser layouts the content. According to the above example, the page is loaded by the layout engine, the description information and layout information of the web page are generated, the description information are compared with the layout information, and thus the method for detecting the hidden content is not affected by the types or encryption, thereby accurately detecting the hidden content and improving accuracy of detecting the hidden content of the web page.

The examples of the present invention also provide a solution for identify a hidden chain.

In the solution, if the hidden content are tag nodes having a function of pointing to a URL and the tag node having the function of pointing to the URL points to an external link, it is determined that the web page is hung with a hidden chain.

The tag node having the function of pointing to the URL may be: a <a> tag node. The introduction of the <a> tag node are as following.

<a> tag may define an anchor, and the anchor generally has the following two usages.

1. establishing a link or a hyperlink pointing to another document, by using a Hypertext Reference (href) attribute.

2. establishing a mark inside a document, by using a name or Identifier attribute, that is, establishing a link pointing to a document segment.

In the following example, the layout engine is the webkit, the description information of the web page is the DOM tree and the layout information of the web page is the render tree, and the solutions are described.

In the procedure of rendering the web page by using the webkit, a dom tree and a render tree are constructed in a memory. The dom tree is used to describe information of a HTML page. The render tree is used to describe the layout of the HTML page, i.e. how to display the dom tree on the screen. In the procedure of constructing the dom tree as the
5 render tree, the webkit determines whether each node of the dom tree is in a hidden state according to a result of parsing the CSS and JavaScript. If the node of the dom tree is in the hidden state, the node is not added to the render tree when constructing the render tree.

By comparing the dom tree and the render tree, many elements included in the dom
10 tree is not contained in the generated render tree. Therefore, by determining whether the nodes included in the dom tree is the same as the nodes included in the render tree, whether the web page has hidden content is determined. Then whether the <a> tag node is included in the hidden content is determined to further determine whether a link is hidden.

15 As shown in Figure 2, the procedure includes the following processing.

At 201, a URL is input.

At 202, the User-Agent of the webkit is configured as a common search engine.

At 203, the URL is visited. Specifically, the URL is loaded by using the webkit, a returned HTML code is parsed, and other elements of a web page, e.g. CSS, JavaScript
20 are loaded. A time period is taken to wait for the execution of the JavaScript, e.g. the time period may be 5 second. The time period may be configured flexibly according to various examples.

At 204, a dom tree and a render tree are generated.

At 205, it is determined whether all HTML tag nodes included in the dom tree is
25 included in the render tree. If all HTML tag nodes included in the dom tree is included in the render tree, processing at 206 is performed. If not all HTML tag nodes included in the dom tree is included in the render tree, it is indicated that the tag node of the dom tree is hidden, and processing at 207 is performed.

At 206, it is determined that the web page is not hung with a hidden chain.

At 207, it is determined whether there are several a tag nodes in the hidden tag nodes of the dom tree. If there are several a tag nodes in the hidden tag nodes of the dom tree, processing at 208 is performed; if there is no a tag node in the hidden nodes of the dom tree, processing at 206 is performed.

At 208, it is determined that the web page is hung with a hidden chain.

It should be noted that, many languages may be used for editing web pages, various languages have different characteristics, and different modes for hiding content are used. In the examples of the present invention, the HTML, CSS, JavaScript are taken as examples and are used to limit the scope of the present invention.

Examples of the present application also provide an apparatus for detecting hidden content of a web page. The apparatus may be located in a terminal device, e.g. a PC, a mobile terminal and etc.

Figure 3 is a schematic diagram illustrating a structure of an apparatus for detecting hidden content of a web page according to various examples of the present invention. As shown in Figure 3, the apparatus includes a loading unit 301, a comparing unit 302 and a content determining unit 303.

The loading unit 301 is to generate description information of a web page and layout information of the web page, by loading the web page.

The comparing unit 302 is to compare the description information of the web page with the layout information of the web page, and determine whether all tag nodes included in the description information of the web page are included in the layout information of the web page.

The content determining unit 303 is to determine at least one tag node included in the description information of the web page and not contained in the layout information of the web page as hidden content, when the comparing unit determines that not all tag nodes included in the description information of the web page are included in the layout information of the web page.

The purpose of configuring hidden content in the web page is not to display content to be hidden on the screen after the browser layouts the content. According to the above example, the page is loaded by the layout engine, the description information and layout information of the web page are generated, the description information are compared with
5 the layout information, and thus the method for detecting the hidden content is not affected by the types or encryption, thereby accurately detecting the hidden content and improving accuracy of detecting the hidden content of the web page.

According to an example, the comparing unit is to, after generating the description information of the web page and the layout information of the web page and a preset
10 delay time period expires, compare the description information of the web page with the layout information of the web page, determine whether all tag nodes included in the description information of the web page are included in the layout information of the web page.

Moreover, as shown in Figure 4, the loading unit 301 is further to load the web page
15 by using the webkit. The apparatus further comprises a loading controlling unit 401.

The loading controlling unit 401 is to, before loading the web page by using the layout engine, configure a user-agent of the layout engine as a search engine.

Many hackers determine whether a web page is visited by a common user or by a search engine crawler according to the user-agent. The hacker only returns a page with
20 SEO to the search engine crawler, and returns normal content to the common user, and thus it is different for network managers to detect abnormal states of the web site. According to the example, when the user-agent of the layout engine is configured as the search engine, the above problem is solved, thereby further improving the accuracy of detecting the hidden content of the web page.

25 According to an example, the loading unit 301 is to load a URL by using the layout engine, parse returned HTML codes, and load other elements in a page corresponding to the returned HTML codes. The other elements include CSS, JavaScript and etc., which is not limited in the example of the present application.

According to an example, the load unit 301 is to load the web page by using a webkit

and generate a dom tree and a render tree.

The comparing unit 302 is to compare the DOM tree with the render tree, determine whether all tag nodes included in the DOM tree are included in the render tree.

5 The content determining unit 303 is to determine at least one tag node included in the DOM tree and not contained in the render tree as the hidden content, when the comparing unit 302 determines that not all tag nodes included in the DOM tree are included in the render tree.

10 It should be noted that, the layout engine is a tool for configuring the layout of the web page and is not limited in webkit according to various examples. Similarly, according to various examples, the description information and the layout information are not limited to the DOM tree and the render tree respectively.

According to an example, a solution for identifying a black link is provided. As shown Figure 5, the apparatus includes a content determining sub-unit 501.

15 The content determining sub-unit 501 is to determining that the web page is hung with a hidden chain, if the hidden content are tag nodes having a function of pointing to a URL and the tag node having the function of pointing to the URL points to an external link.

It should be noted that, in the above examples, the units are divided according logic functions.

20 The methods and modules described herein may be implemented by hardware, machine-readable instructions or a combination of hardware and machine-readable instructions. Machine-readable instructions used in the examples disclosed herein may be stored in storage medium readable by multiple processors, such as hard drive, CD-ROM, DVD, compact disk, floppy disk, magnetic tape drive, RAM, ROM or other proper storage device. Or, at least part of the machine-readable instructions may be substituted
25 by specific-purpose hardware, such as custom integrated circuits, gate array, FPGA, PLD and specific-purpose computers and so on.

A machine-readable storage medium is also provided, which is to store instructions

to cause a machine to execute a method as described herein. Specifically, a system or apparatus having a storage medium that stores machine-readable program codes for implementing functions of any of the above examples and that may make the system or the apparatus (or CPU or MPU) read and execute the program codes stored in the storage
5 medium.

In this situation, the program codes read from the storage medium may implement any one of the above examples, thus the program codes and the storage medium storing the program codes are part of the technical scheme.

The storage medium for providing the program codes may include floppy disk, hard
10 drive, magneto-optical disk, compact disk (such as CD-ROM, CD-R, CD-RW, DVD-ROM, DVD-RAM, DVD-RW, DVD+RW), magnetic tape drive, Flash card, ROM and so on. Optionally, the program code may be downloaded from a server computer via a communication network.

It should be noted that, alternatively to the program codes being executed by a
15 computer, at least part of the operations performed by the program codes may be implemented by an operation system running in a computer following instructions based on the program codes to realize a technical scheme of any of the above examples.

In addition, the program codes implemented from a storage medium are written in storage in an extension board inserted in the computer or in storage in an extension unit
20 connected to the computer. In this example, a CPU in the extension board or the extension unit executes at least part of the operations according to the instructions based on the program codes to realize a technical scheme of any of the above examples.

The foregoing is only preferred examples of the present invention and is not used to limit the protection scope of the present invention. Any modification, equivalent
25 substitution and improvement without departing from the spirit and principle of the present invention are within the protection scope of the present invention.

Claims

1. A method for detecting hidden content of a web page, comprising:
loading a web page;
generating description information of the web page and generating layout
5 information of the web page;
comparing the description information of the web page with the layout information
of the web page according to a preconfigured rule;
determining whether all tag nodes comprised in the description information of the
web page are contained in the layout information of the web page; and
10 determining at least one tag node, which being comprised in the description
information of the web page and not being contained in the layout information of the web
page, as hidden content.
2. The method of claim 1, wherein comparing the description information of the web
page with the layout information of the web page comprises:
15 after a preset delay time period expires, comparing the description information of the
web page with the layout information of the web page.
3. The method of claim 1, wherein loading the web page comprises:
configuring a user-agent character string of an layout engine as a search engine;
loading the web page by using the layout engine.
- 20 4. The method of claim 3, wherein loading the web page by using the layout engine
comprises:
loading a Uniform/Universal Resource Locator (URL) by using the layout engine;
parsing returned HTML codes; and loading other elements in a page corresponding to the
returned HTML codes.
- 25 5. The method of claims 3 or 4, wherein
the layout engine is a is an open-source browser layout engine; the description
information of the web page is a Document Object Model (DOM) tree; and the layout
information of the web page is a render tree.
6. The method of any of claims 1 to 4, further comprising:
30 determining whether the web page is hung with a hidden chain according to the
hidden content.

7. The method of claim 6, wherein determining whether the web page comprises a hidden chain according to the hidden content comprises:

if the hidden content is a tag node having a function of pointing to a URL, and the tag node having the function of pointing to the URL points to an external link,
5 determining that the web page is hung with the hidden chain.

8. An apparatus for detecting hidden content of a web page, comprising:

a loading unit, configured to load a web page and to generate description information of a web page and generate layout information of the web page;

a comparing unit, configured to compare the description information of the web page
10 with the layout information of the web page according to a preconfigured rule, and further to determine whether all tag nodes comprised in the description information of the web page are contained in the layout information of the web page; and

a content determining unit, configured to determine at least one tag node, which being comprised in the description information of the web page and not being contained
15 in the layout information of the web page, as hidden content.

9. The apparatus of claim 8, wherein

the comparing unit is to, after a preset delay time period expires, compare the description information of the web page with the layout information of the web page, determine whether all tag nodes comprised in the description information of the web page
20 are comprised in the layout information of the web page.

10. The apparatus of claim 8, further comprising:

a loading controlling unit, to configure a user-agent character string of an layout engine as a search engine;

wherein the loading unit is to loading the web page by using a layout engine.

25 11. The apparatus of claim 10, wherein

the loading unit is to load a Uniform/Universal Resource Locator (URL) by using the layout engine; parse returned HTML codes; and load other elements in a page corresponding to the returned HTML code.

12. The apparatus of claims 10 or 11, wherein

30 the loading unit is to loading the web page by using an open-source browser layout engine, generate a Document Object Model (DOM) tree and a render tree;

the comparing unit is to compare the DOM tree with the render tree, determine whether all tag nodes comprised in the DOM tree are comprised in the render tree;

a content determining unit, to determine at least one tag node comprised in the DOM tree and not contained in the render tree as hidden content, when the comparing unit determines that not all tag nodes comprised in the DOM tree are comprised in the render tree.

5 13. The apparatus of any of claims 8 to 11, further comprising:

a content determining sub-unit, to determining whether the web page is hung with a hidden chain according to the hidden content.

14. The apparatus of claim 13, wherein

10 the content determining sub-unit is to determine that the web page is hung with the hidden chain, if the hidden content is a tag node having a function of pointing to a URL and the tag node having the function of pointing to the URL points to an external link.

1/2

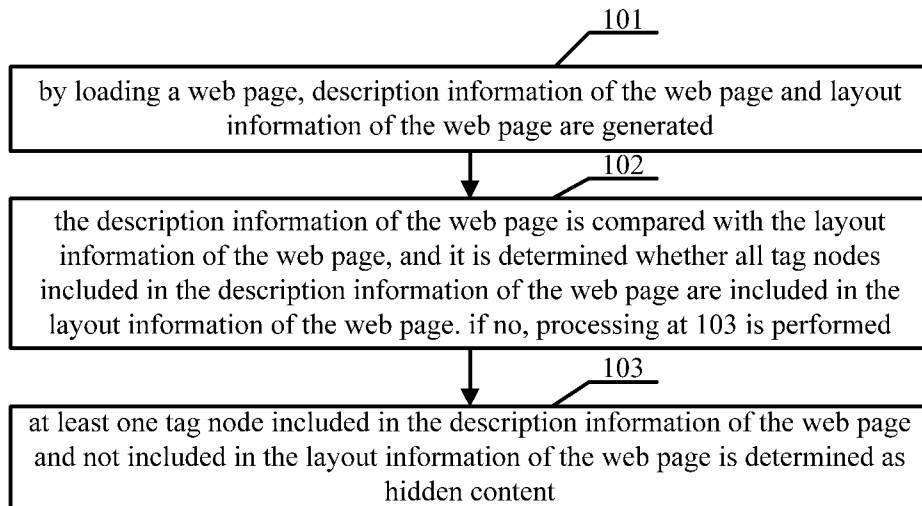


Fig. 1

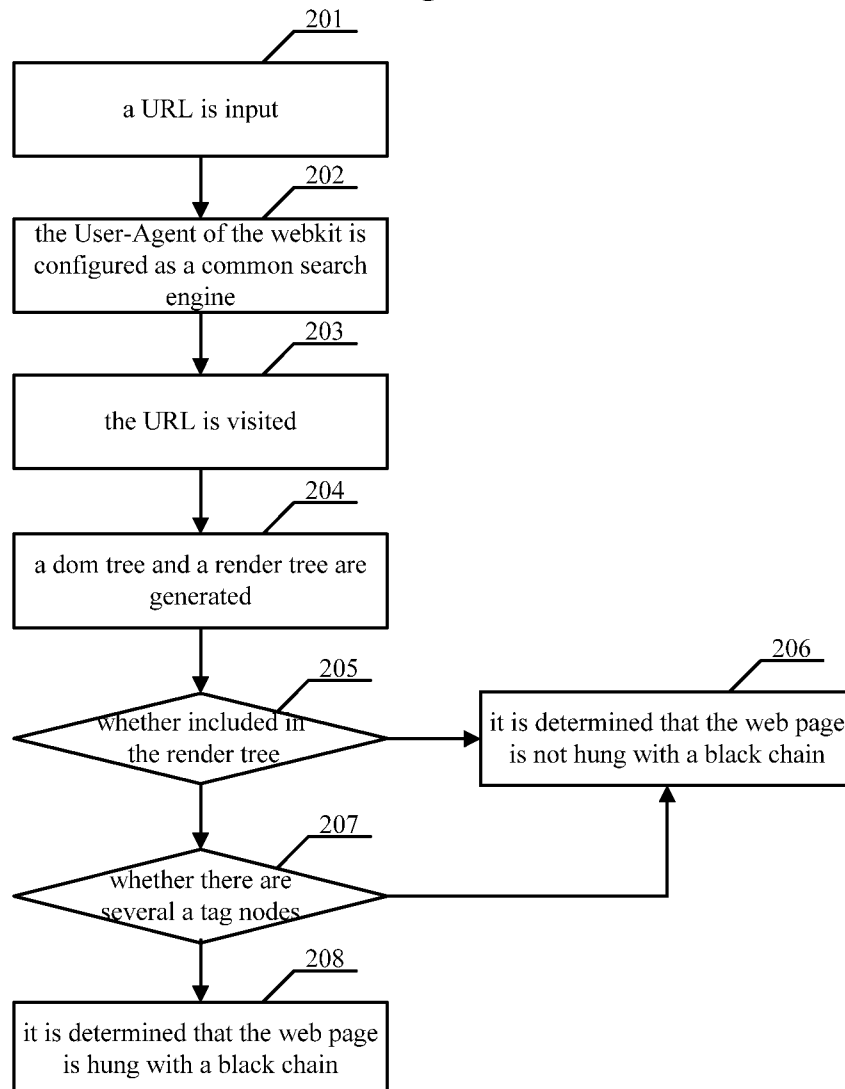


Fig. 2

2/2

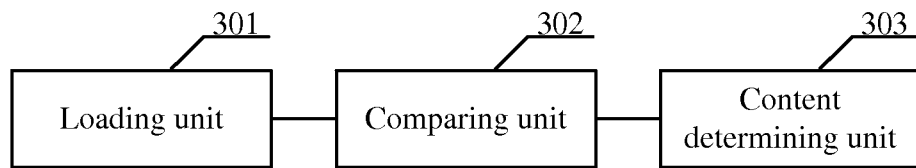


Fig. 3

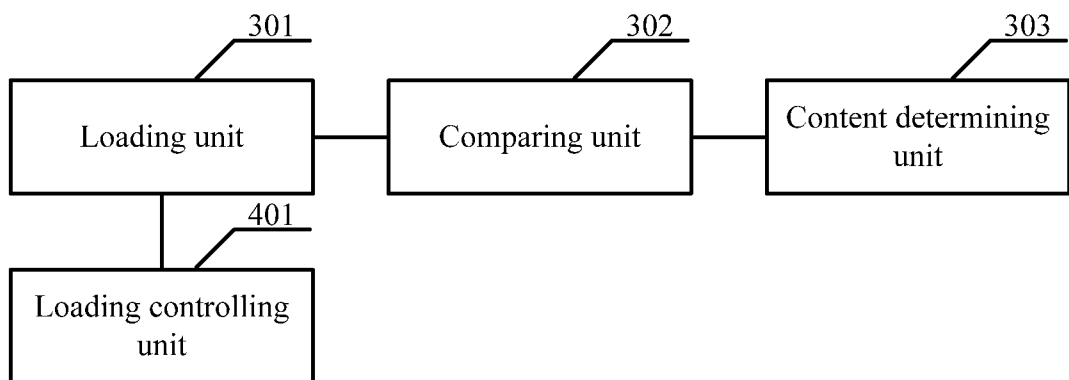


Fig. 4

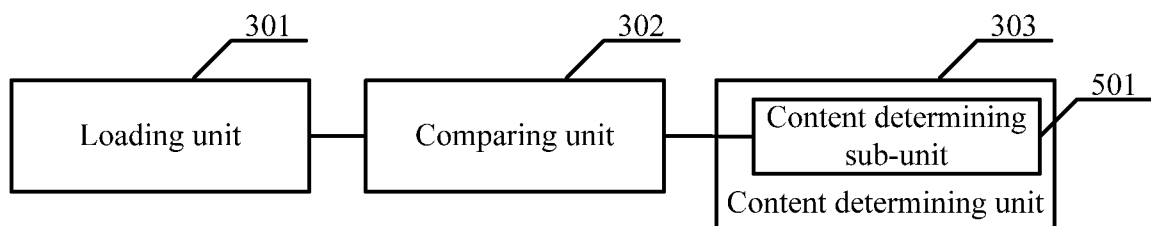


Fig. 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2013/085367

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30 (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, G06K

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, CNTXT, CNKI, DWPI, EPODOC, IEEE: webpage, hidden, conceal, ensconce, secret, shroud, determine, identify, Malicious, black ,compare, layout, URL, render

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2010114902 A1 (Brigham young university) 06 May 2010(06. 05. 2010) see claims 1-20 and figures 2-5	1, 8
Y	WO 2009049275 A1 (Symantec corporation) 16 April 2009 (16. 04. 2009) see claims 1-20 and figures 1, 2	1, 8
A	US 2012255027 A1 (INFOSYS TECHNOLOGIES LTD) 04 October 2012 (04. 10. 2012) see the whole document	1-14

☒ Further documents are listed in the continuation of Box C.☒ See patent family annex.

* Special categories of cited documents:	“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
“A” document defining the general state of the art which is not considered to be of particular relevance	“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
“E” earlier application or patent but published on or after the international filing date	“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
“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)	“&” document member of the same patent family
“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	

Date of the actual completion of the international search
06 January 2014 (06. 01. 2014)Date of mailing of the international search report
16 Jan. 2014 (16.01.2014)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. 86-10-62019451Authorized officer
xubo
Telephone No. (86-10)62411704

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2013/085367

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 102542201 A (BEIJING QIHOO SCI&TECHNOLOGY CO LTD) 04 July 2012 (04. 07. 2012) see the whole document	1-14
A	CN 102682097 A (BEIJING NSFOCUS INFORMATION SECURITY TEC) 19 September 2012 (19. 09. 2012) see the whole document	1-14
A	CN 102622543 A (BEIJING BAIDU NETCOM SCI&TECHNOLOGY CO) 01 August 2012 (01. 08. 2012) see the whole document	1-14

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/CN2013/085367

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
US 2010114902 A1	06. 05. 2010	None	
WO 2009049275 A1	16. 04. 2009	None	
US 2012255027 A1	04. 10. 2012	IN 201101051 I4	19. 10. 2012
CN 102542201 A	04. 07. 2012	None	
CN 102682097 A	19. 09. 2012	None	
CN 102622543 A	01. 08. 2012	None	