



- (51) **International Patent Classification:**
G06F 17/30 (2006.01)
- (21) **International Application Number:**
PCT/CN2014/083261
- (22) **International Filing Date:**
29 July 2014 (29.07.2014)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (30) **Priority Data:**
201310326576.2 30 July 2013 (30.07.2013) 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: LUO, Yan**; Room 403, East Block 2, SEG Park, Zhenxing Road, Futian District, Shenzhen, Guangdong 518000 (CN).
- (74) **Agent: BEIJING SAN GAO YONG XIN INTELLECTUAL PROPERTY AGENCY CO., LTD.**; 703, Grand Place, No.5 HUIZHONG ROAD, Chaoyang District, Beijing 100101 (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 CLUSTERING PHISHING WEBPAGES

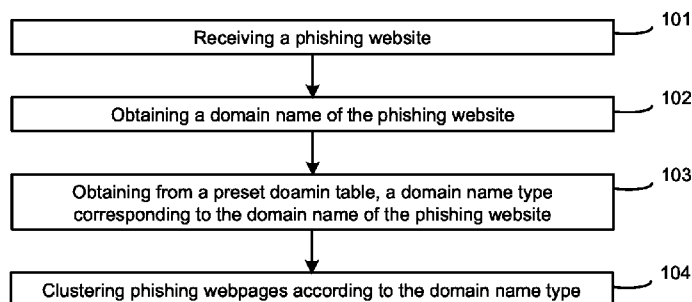


Figure 1

(57) **Abstract:** The present disclosure discloses a method and device for clustering phishing webpages. The method includes: receiving by a device, a phishing website; obtaining a domain name of the phishing website; obtaining from a preset domain name table, a domain name type corresponding to the domain name of the phishing website; and clustering phishing webpages according to the domain name type. In this regard, when phishers use secondary level domain (SLD) names for phishing, the present disclosure may sort out and cluster the phishing webpage based on the domain name type within the domain names in the SLD, thereby reducing the false-positive rate of phishing webpages and thoroughly include the phishing webpages spread from the source.

METHOD AND DEVICE FOR CLUSTERING PHISHING WEBPAGES

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] The application claims priority to Chinese Patent Application No. 2013103265762, filed on July 30, 2013, which is incorporated by reference in its entirety.

FIELD OF THE TECHNOLOGY

[0002] The present disclosure specifically relates to a method and a device for clustering phishing webpages in the field of information security.

BACKGROUND

[0003] A phishing webpage is a disguised bank webpage or e-commerce webpage that may steal private information such as bank accounts and passwords submitted by users. The so-called phishing webpage is essentially an online fraud, in which phishers counterfeit the URLs and the page content of real websites by different means, or insert dangerous HTML codes into certain webpages of real websites by taking advantage of loopholes in the server programs of the real websites, so as to trick people into divulging their private information, for example, accounts and passwords of bank/credit cards. Clustering of phishing webpages is to cluster the phishing-purposed webpages for use as the criteria for comparison in detecting phishing webpages.

[0004] In the prior art, there are diversified methods for clustering phishing webpages. Conventional methods for clustering phishing webpages usually include: firstly, determining a standard time range, for example, a normal day; secondly, presetting a threshold value, and obtaining the number of detected phishing webpages in any site or domain; thirdly, determining whether the obtained quantity of detected phishing webpages exceeds the preset threshold value, and label the entire sites or entire domains associated with the number of detected phishing webpages that exceed the preset threshold value as phishing webpages.

[0005] For better understanding of domain names, domain names are substitutes for IP addresses, which are presented numerically to further simplify the process of searching for information on the Internet. Domain names servers maintain a list that matches domain names to IP addresses. There are two types of domain names: (A) Top Level Domains (TLD) having domain names such as: .com, .edu, .gov, .net, .org, etc., and (B) secondary level domains (SLD) having domain names which precede TLDs in the web addresses, and usually refer to the

organizations (e.g., msn.com, uspto.gov, edu.net, lsm.org, etc.) that have been registered to the TLD through Internet Corporation for Assigned Names and Numbers (ICANN).

[0006] Therefore, when phishers utilizes secondary level domain names of SLDs for phishing, the prior art methods usually merely cluster phishing webpages into sites or domains to label an entire SLDs as being phishing webpages. Such practice inadvertently also include many other innocent secondary level domain names under the SLDs as also being phishing webpages, thus may incorrectly and unfairly be also labeled as the offending phishing webpages.

[0007] For example, a phisher may request to use a secondary level domain name (e.g., a.cn.ms) for phishing activities. The prior art methods of clustering phishing webpages may label a large number of secondary level domain (SLD) names in the entire SLD of cn.ms, namely, including not only the sub-domain name of a.cn.ms (i.e., the offending phishing webpage), but also other sub-domain names such as b.cn.ms to ...n.cn.ms which are registered under the SLD of cn.ms. In this regard, the innocent sub-domain names (e.g., b.cn.ms to ...n.cn.ms) under the SLD of cn.ms not only are being mistakenly labeled as phishing webpages, but also in effect, help to hide the offending phishing webpages which are originated from the sub-domain name a.cb.ms as the source. In addition, the labeling of the other innocent sub-domain names (e.g., b.cn.ms to ...n.cn.ms) has also caused an erroneously high false positive rate in the clustering of phishing webpages.

[0008] In another example, assuming that only three sub-domain names: b.a.cn.ms, c.a.cn.ms, and d.e.a.cn.ms of a.cn.ms may be the phishing webpages. Sophisticated phishers may exploit “extensive domain name analysis” to generate a large number of sub-domain names “*.a.cn.ms” (i.e., sub-domain names of a.cn.ms) within an extremely short period. The prior art methods for clustering phishing webpages may include the entire site or every sub-domain names (i.e., “*.a.cn.ms”) under the sub-domain of a.cn.ms as phishing webpages. As seen, the prior art methods are ineffective in precisely clustering only the source of the phishing webpages (i.e., b.a.cn.ms, c.a.cn.ms, and d.e.a.cn.ms).

SUMMARY

[0009] The present disclosure provides a method and device for clustering phishing webpages, which reduces the false positive rate of phishing webpages and improve effectiveness in precisely clustering the phishing webpages which are originated from the source.

[0010] An embodiment of the disclosure provides a method for clustering phishing webpages,

including the following operations of: receiving by the terminal, a phishing website; obtaining a domain name of the phishing website; obtaining from a preset domain name table, a domain name type corresponding to the domain name of the phishing website; and clustering phishing webpages according to the domain name type.

[0011] Preferably, the clustering of phishing webpages according to the domain name type further includes: determining whether the domain name type pertains to a secondary level domain (SLD) name, and if yes, obtaining a sub-domain of the domain name of the phishing website; when a preset cluster information library does not include the sub-domain, obtaining a count result of the sub-domain after incrementing by 1; and determining whether the count result of the sub-domain meets clustering requirements, and if yes, clustering the sub-domain of the domain name of the phishing website into the cluster information library.

[0012] Preferably, the method further includes: when the domain name type does not pertain to the SLD name, **obtaining a count result of the domain name of the phishing website after incrementing by 1**; determining whether the count result of the domain name of the phishing website satisfies clustering requirements, and if yes, clustering the domain name of the phishing website into the cluster information library.

[0013] Preferably, the clustering requirements include one of: the count result of the sub-domain exceeds a first preset threshold value within a first preset time; or a ratio of the count result of the sub-domain to websites of an entire domain or to websites of the sub-domain exceeds a first preset ratio within the first preset time.

[0014] Alternately, the clustering requirements may include one of: the count result of the domain name of the phishing website exceeds a second preset threshold value within a second preset time; or a ratio of the domain name of the phishing website to websites of an entire domain or to websites of the sub-domain exceeds a second preset ratio within the second preset time.

[0015] Another embodiment of the disclosure discloses a device for clustering phishing webpages. The device may include: at least a processor with circuitry operating in conjunction with at least a memory storing codes to be executed to perform functions as a plurality of modules or sub-modules, wherein the plurality of modules or sub-modules include: a receiving module, which causes the device to receive a phishing website; a first acquisition module, which causes the device to obtain the domain name of the phishing website; a second acquisition module, which causes the device to obtain from a preset domain name table, a domain name type corresponding to the domain name of the phishing website; and a clustering

module, which causes the device to cluster phishing webpages according to the domain name type.

[0016] Preferably, the clustering module includes: a first determining sub-module, which causes the device to determine whether the domain name type pertains to a secondary level domain (SLD) name; a first acquisition sub-module, which causes the device to obtain a sub-domain of the domain name of the phishing website when it is determined that the domain name type pertains to the secondary level domain (SLD) name; a first addition sub-module, which causes the device **to obtain a count result of the sub-domain after incrementing by 1**, when a preset cluster information library does not include the sub-domain; a second determining sub-module, which causes the device to determine whether the count result of the sub-domain meets clustering requirements; and a first clustering sub-module, which causes the device to cluster the sub-domain of the domain name of the phishing website into the cluster information library when it is determined that the count result of the sub-domain meets clustering requirements.

[0017] Preferably, the clustering module includes: a second addition sub-module, which causes the device **to obtain a count result of the domain name of the phishing website after incrementing by 1**, when it is determined that the domain name type does not pertain to the SLD name; a third determining sub-module, which causes the device to determine whether the count result of the domain name of the phishing website satisfies clustering requirements; and a second clustering sub-module, which causes the device to cluster the domain name of the phishing website into the cluster information library when it is determined that the count result of the domain name of the phishing website satisfies clustering requirements.

[0018] Compared with the prior art clustering phishing webpages method, the present disclosure is capable of clustering phishing webpages according to the domain name type upon obtaining the domain name type corresponding to a phishing website. In this regard, even when a phisher using sub-domain names of SLDs for phishing to hide its source, the present disclosure may still be able to precisely cluster the phishing webpages which are originated from the source, thus reduce the false-positive rate in detecting phishing webpages without inclusion of legitimate websites by mistake.

BRIEF DESCRIPTION OF THE DRAWINGS

[0019] The accompanying drawings may be included to provide further understanding of the

claims and disclosure which may be incorporated in, and constitute a part of this specification. The detailed description and illustrated embodiments described may serve to explain the principles defined by the claims.

[0020] Figure 1 shows the flowchart illustrating a method for clustering phishing webpages, according to an embodiment of the disclosure;

[0021] Figure 2 shows a distribution diagram of different domain name types of phishing websites;

[0022] Figure 3 shows a flowchart of a method for clustering phishing webpages, according to another embodiment of the disclosure;

[0023] Figure 4 shows an exemplary structure diagram of a device for clustering phishing webpages, according to an embodiment of the disclosure;

[0024] Figure 5 shows an exemplary structure diagram of a server provided in Figure 4.

DETAILED DESCRIPTION

[0025] The various embodiments of the disclosure may be further described in details in combination with attached drawings and embodiments below. It should be understood that the specific embodiments described here may be used only to explain the disclosure, and may not be used to limit the disclosure. In addition, for the sake of keeping description brief and concise, the newly added features, or features which may be different from those previously described in each new embodiment will be described in details. Similar features may be referenced back to the prior descriptions in a prior numbered drawing or referenced ahead to a higher numbered drawing. Unless otherwise specified, all technical and scientific terms herein may have the same meanings as understood by a person skilled in the art.

[0026] Embodiment 1

[0027] There are a diversity of methods for clustering phishing webpages in the prior art, but none of them utilizes the feature of “domain name type” to cluster the phishing webpages which are originated from the source without inclusion of other legitimate websites by mistake. The embodiments in the disclosure solve the described problems of clustering only the phishing webpages from the source and avoiding falsely creating a high positive rate in detecting phishing webpages.

[0028] Domain name types may include a top-level domain (TLD) and a secondary level

domain (SLD). For example, “.com” is a TLD which may allow a user to request its sub-domains; where “cn.ms” is a SLD which may allow a user to register sub-domain names which are lower levels under the SLD (e.g., a.cn.ms, b.cn.ms... n.cn.ms, and b.a.cn.ms...n.n.cn.ms, d.b.e.a.cn.ms, to name a few, etc.).

[0029] When phishers use sub-domain names of SLDs for phishing, the methods in the prior art for clustering phishing webpages have two disadvantages: namely, unable to precisely cluster only the offending phishing webpages without inclusion of other innocent webpages under the same domain name, and as a result, a high false positive rate for labeling phishing webpages. The present disclosure utilizes the feature of “domain name type” in the method and device embodiments to overcome the disadvantages of the prior art methods.

[0030] Figure 1 shows a flowchart illustrating a method for clustering phishing webpages, according to an embodiment of the disclosure. The method may include at least the following operations:

[0031] Step 101: Receiving a phishing website. A phishing website may be received irrespective of the detection method.

[0032] Step 102: Obtaining a domain name of the phishing website. The obtained domain name of the phishing website may be a text address associated with a numerical IP address. For example: b.a.cn.ms may be a phishing website under a domain name of cn.ms, and b.a.com/1.asp may be a phishing website under a domain name of a.com.

[0033] Step 103: Obtaining from a preset domain name table, a domain name type corresponding to the domain name of the phishing website. Upon obtaining the domain name, a preset domain name table may be queried to look for a domain name type corresponding to the obtained domain name, which the domain name type may either be a SLD type or non-SLD type.

[0034] Table 1 may illustrate an exemplary preset domain name table. The format of a specific domain name table may not be limited to that of Table 1. In addition, the domain name table may be obtained with the results from artificial statistics. Table 1 may display domain name types of different domain names as follows:

Domain Name Type	Domain Name
1	tk
1	co.cc
2	in
2	info

3	com
4	cn
5	cn.ms
5	net.tf
6	3322.org
7	vicp.net

Table 1

[0035]

In an embodiment, the types of domain name may be classified according to domain name fees structure. More specifically, classifying domain name types based on domain name fees may result in at least two domain name types: namely, those of free domains (including free TLDs and free SLDs) and those of low cost domain names (including low cost TLDs and low cost SLDs, such as dynamic domains).

[0036] It may be pointed out that free or low cost domain names are more likely being utilized by phishers due to low expenses for phishing. Since low cost or free domain names are increasingly becoming a phish-stricken area, in this regard, classifying domain name types according to domain name fees and cost may become a factor in determining whether the domain name type may be a phishing webpage or not.

[0037] Figure 2 shows an exemplary distribution diagram of domain names which are infested with phishing websites domain name types (data source: 2012 Annual Report of Anti-Phishing Alliance of China). It may be seen that the domain names excluding .com are basically free domains or low cost domains. .tk, .co.cc, and .pl are representative free TLDs which account for a large ratio. Under the TLDs of .ms and .tf, they include a large number of free SLDs including cn.ms, hk.ms, net.tf, and eu.tf. Low cost TLDs may include: .to, .info, and .in, to name a few. Low cost SLDs in China are mostly provided by dynamic domain name providers, such as 3322.org.

[0038] Step 104: Clustering phishing webpages according to the domain name type. More specifically, a domain name may correspond to a domain name type. The domain name type of the domain name may be determined to one which pertains to the SLD, and clustering of phishing webpages may be performed according to the determination result.

[0039] To summarize, the present disclosure firstly receives any phishing website; secondly, obtains a domain name of the phishing website; thirdly, obtains a domain name type corresponding to the domain name from a preset domain name table; finally, clusters

phishing webpages according to the domain name type. In this regard, when phishers use secondary level domain names of SLDs for phishing, the present disclosure may sort out and cluster the phishing webpage based on the domain name type within the domain names in the SLD, thereby reducing the false-positive rate of phishing webpages and thoroughly include the phishing webpages spread from the source.

[0040] Embodiment 2

[0041] Figure 3 shows a flowchart of the method for clustering phishing webpages, according to another embodiment. The method may include the following operations:

[0042] Step 301: Receiving any phishing website;

[0043] Step 302: Obtaining the domain name of the phishing website;

[0044] Step 303: Obtaining from a preset domain name table, the domain name type corresponding to the domain name;

[0045] In this embodiment, Steps 301 to 303 are the same as Steps 101 to 103 in the embodiment of Fig. 1.

[0046] Step 304: Determining whether the domain name type is of SLD name, if yes, proceeding to Step 305, if not, proceeds to Step 309. More specifically, the domain name type corresponding to the domain name may be determined by referencing to Table 2. Table 2 is an exemplary domain name type table, which may be obtained by artificial statistics. Table 2 may be further developed from Table 1, and is not limiting in its content.

[0047] Specifically, Table 2 may be as follows:

SLD or not	Domain Name Type	Name of Domain Name Type	Threshold of Number of Daily Blacklisted Entire Domain Websites	Threshold of Ratio of Number of Daily Clustered Websites
0	1	Free TLD	50	50%
0	2	Low cost TLD	70	60%
0	3	General TLD	100	80%
0	4	Trusted TLD	200	90%
1	5	Free SLD	30	30%
1	6	Low cost SLD	60	50%
1	7	General SLD	100	70%

Table 2

[0048] Step 305: Obtaining a sub-domain of the domain name of the phishing website. In this embodiment, when the domain name type pertains to a SLD name, the sub-domain name of the phishing website's domain name may be obtained. For example, if a phishing website is b.a.cn.ms, then its SLD name may be cn.ms with at least a sub-domain name of a.cn.ms. Various methods may be used to obtain a sub-domain name from a SLD name, which is non-restricting to the disclosure.

[0049] Step 306: When the preset cluster information library does not include the sub-domain name, obtaining a count result of the sub-domain after incrementing by 1. More specifically, a real-time count of the detected sub-domains may be performed, and a sub-domain count may be incremented by 1 each time a sub-domain is detected, wherein, the sub-domain counting method is not limited by the disclosure of the embodiment.

[0050] Step 307: Determining whether the count result of the sub-domain meets clustering requirements, and if yes, proceeding to Step 308, if not, proceed to cluster other phishing webpages. More specifically, the determination is performed upon obtaining the result of the sub-domain count. In practice, the clustering requirement may include a preset threshold value and within a preset time. If the count result of the sub-domain is larger than the preset threshold value within the preset time, or the ratio of the count result to the websites of the entire domain or to the websites of the sub-domain is larger than a preset ratio within the preset time.

[0051] Referring to Table 2, the present disclosure may set the clustering requirement to "The

threshold of number of daily blacklisted entire domain websites is 50”, then, when the daily secondary level domain name count is larger than 50, proceeds to Step 308. Likewise, the clustering requirement may be set to “The threshold of ratio of daily clustered websites being 50%”, then when the ratio of the daily secondary level domain count result to the number of all blacklisted websites is larger than 50%, proceeds to Step 308.

[0052] Step 308: Clustering the sub-domain name of the domain name into the cluster information library. More specifically, when the result of the sub-domain count meets the preset clustering requirements, clusters the sub-domain into the cluster information library and labels the sub-domain as a phishing webpage.

[0053] Step 309: **Obtaining a count result of the domain name of the phishing website after incrementing by 1.** More specifically, when the domain name type is not of SLD name, the domain name count is incremented by 1 to obtain the count result of the domain name of the phishing website.

[0054] Step 310: Determining whether the count result of the domain name satisfies the clustering requirements, and if yes, proceeding to Step 311, if not, proceeds to cluster other phishing webpages.

[0055] More specifically, the preset clustering requirement may be obtained by querying Table 2. For example, for the domain name with the domain name type “1”, a corresponding clustering requirement may be set to “The threshold of number of daily blacklisted websites in the entire domain being 50”. In other words, the number of daily clustered websites in the entire domain is larger than 50.

[0056] Step 311: Clustering the domain name into the cluster information library, after the domain name count meeting the preset clustering requirements. Clustering of phishing webpages may be according to the domain name type, even when phishers use secondary level domain names of SLDs for phishing. Accordingly, this embodiment may effectively reduce the false-positive rate in detecting phishing webpages and thoroughly include the spread of phishing webpages at the source.

[0057] Embodiment 3

[0058] Figure 4 shows an exemplary structure diagram of a device (400) for clustering phishing webpages, according to an embodiment of the disclosure. The device may include: at least a processor with circuitry (407) operating in conjunction with at least a memory (405)

storing codes to be executed to perform functions as a plurality of modules or sub-modules, wherein the plurality of modules or sub-modules may include at least a receiving module (401), a first acquisition module (402), a second acquisition module (403) and a clustering module (404).

[0059] The receiving module (401) causes the device (400) to receive a phishing website. The first acquisition module (402) causes the device (400) to obtain a domain name of the phishing website. The second acquisition module (403) causes the device (400) to obtain from a preset domain name table, a domain name type corresponding to the domain name of the phishing website; and the clustering module (404) causes the device (400) to cluster phishing webpages according to the domain name type.

[0060] Wherein, the clustering module (404) may further include:

[0061] a first determining sub-module, which causes the device to determine whether the domain name type pertains to a SLD;

[0062] a first acquisition sub-module, which causes the device to obtain a sub-domain of the domain name of the phishing website when the determination result of the first determining sub-module is yes (i.e., the domain name type pertains to the SLD name);

[0063] a first addition sub-module, which causes the device to increment by 1 to obtain a count result of the sub-domain, when the preset cluster information library does not include the sub-domain;

[0064] a second determining sub-module, which causes the device to determine whether the count result of the sub-domain meets the clustering requirements;

[0065] a first clustering sub-module, which causes the device to cluster the sub-domain of the domain name of the phishing website into the cluster information library when the determination result of the second determining sub-module is yes (i.e., the count result of the sub-domain meets clustering requirements).

[0066] The clustering module may further include: a second addition sub-module, which causes the device to increment by 1 to obtain a count result of the domain name of the phishing website when it is determined that the domain name type does not pertain to the SLD name; a third determining sub-module, which causes the device to determine whether the count result of the domain name of the phishing website satisfies clustering requirements; and a second clustering sub-module, which causes the device to cluster the domain name of the

phishing website into the cluster information library when it is determined that the count result of the domain name of the phishing website satisfies clustering requirements.

[0067] Figure 5 illustrates a server provided in this embodiment, wherein the server (500) may cause the device (400) to carry out the method provided in the foregoing embodiments described in Figs. 1-2. Specifically: the server (500) may include one or more memories (510) made of non-transitory readable storage media, an input unit (520), an output unit (530), a processor (540) comprising one or more processing cores, and a power supply (550).

[0068] The memory (510) may cause the device to store software programs and modules. By running a software program or module stored in the memory (510), the processor (540) may execute various functional applications and data processing. The memory (510) may mainly function as a storage program area and a storage data area, wherein the storage program area may store an operating system, application programs required for at least one function, and the storage data area may store the data based on usage of the computer. In addition, the memory (510) may be such as a high-speed random access memory (RAM) and a non-volatile memory, for example, at least one disk storage device, a flash memory device, or a volatile solid memory device of another type. Correspondingly, the memory (510) may include a memory controller, which enables the processor (540) and the input unit (520) to access the memory (510).

[0069] The input unit (520) may be one which causes the device (400) to receive the inputted numerical or character information, and generate keyboard inputs, mouse inputs, operating rod inputs, optical inputs, or trackball signal inputs related to user setting and functional control.

[0070] As the control center of the server (500), the processor (540) connects different parts of the server by means of various interfaces and lines, and performs various functions of the computer and processes data by running or executing the software program and/or module stored in the memory (510) and invoking the data stored in the memory (510), thereby performing overall monitoring to the mobile phone. Optionally, the processor (540) may include one or more processing cores.

[0071] The server also comprises the power supply (550) (such as a battery) that supplies power to various parts of the server. Preferably, the power supply (350) may be connected to the processor (540) logically through a power management system, so that the power management system may perform such functions as charging management, discharging management, and power consumption management. The power supply (550) may further comprise one or

multiple DC/AC power supplies, a recharging system, a power failure detection circuit, a power converter or inverter, and a power status indicator.

[0072] In this embodiment, the processor (540) may specifically load the executable files associated with the processes of one or multiple applications to the memory (510) according to the following instructions, and run the applications stored in the memory (510), thereby implementing various clustering phishing webpage functions corresponding to Figs. 1-4. For example, the server may perform functions including: receiving by a device, a phishing website; obtaining a domain name of the phishing website; obtaining from a preset domain name table, a domain name type corresponding to the domain name of the phishing website; and clustering phishing webpages according to the domain name type.

[0073] The modules and units described as separate parts may be or may not be physically separated, and parts displayed as units may be or may not be physical units, that is, they may be located at the same place or be distributed on multiple network units. Some or all of the modules may be selected as needed to attain the objectives of this embodiment. A person with ordinary skills in the art may understand and carry out it without inputting creative effort.

[0074] The sequence numbers of the above-mentioned embodiments may be intended only for description, instead of indicating the relative merits of the embodiments. It should be understood by those with ordinary skill in the art that all or some of the steps of the foregoing embodiments may be implemented by hardware, or software program codes stored on a non-transitory computer-readable storage medium with computer-executable commands stored within. For example, the disclosure may be implemented as an algorithm as codes stored in a program module or a system with multi-program-modules. The computer-readable storage medium may be, for example, nonvolatile memory such as compact disc, hard drive, ROM or flash memory. The computer-executable commands may enable a computer, a server, a smart phone, a tablet or any similar computing device to render clustering of phishing webpages operations.

Claims

1. A method for clustering phishing webpages in a device, comprising:
receiving a phishing website;
obtaining a domain name of the phishing website;
obtaining from a preset domain name table, a domain name type corresponding to the domain name of the phishing website ; and
clustering phishing webpages according to the domain name type.
2. The method according to claim 1, wherein the clustering of phishing webpages according to the domain name type comprises:
determining whether the domain name type pertains to a secondary level domain (SLD) name, and if yes, obtaining a sub-domain of the domain name of the phishing website;
when a preset cluster information library does not include the sub-domain, obtaining a count result of the sub-domain after incrementing by 1; and
determining whether the count result of the sub-domain meets clustering requirements, and if yes, clustering the sub-domain of the domain name of the phishing website into the cluster information library.
3. The method according to claim 2, wherein the method further comprises:
when the domain name type does not pertain to the SLD name, obtaining a count result of the domain name of the phishing website after incrementing by 1; and
determining whether the count result of the domain name of the phishing website satisfies clustering requirements, and if yes, clustering the domain name of the phishing website into the cluster information library.
4. The method according to claim 2, wherein the clustering requirements comprises one of:
the count result of the sub-domain exceeds a first preset threshold value within a first preset time; or
a ratio of the count result of the sub-domain to websites of an entire domain or to websites

of the sub-domain exceeds a first preset ratio within the first preset time.

5. The method according to claim 3, wherein the clustering requirements comprises one of:
the count result of the domain name of the phishing website exceeds a second preset threshold value within a second preset time; or

a ratio of the domain name of the phishing website to websites of an entire domain or to websites of the sub-domain exceeds a second preset ratio within the second preset time.

6. A device for clustering phishing webpages, comprises at least a processor with circuitry operating in conjunction with at least a memory storing codes to be executed to perform functions as a plurality of modules or sub-modules, wherein the plurality of modules or sub-modules comprise:

a receiving module, which causes the device to receive a phishing website;

a first acquisition module, which causes the device to obtain a domain name of the phishing website;

a second acquisition module, which causes the device to obtain from a preset domain name table, a domain name type corresponding to the domain name of the phishing website; and

a clustering module, which causes the device to cluster phishing webpages according to the domain name type.

7. The device according to claim 6, wherein the clustering module comprises:

a first determining sub-module, which causes the device to determine whether the domain name type pertains to a secondary level domain (SLD) name;

a first acquisition sub-module, which causes the device to obtain a sub-domain of the domain name of the phishing website when it is determined that the domain name type pertains to the secondary level domain (SLD) name;

a first addition sub-module, which causes the device to obtain a count result of the sub-domain after incrementing by 1, when a preset cluster information library does not include the sub-domain;

a second determining sub-module, which causes the device to determine whether the count result of the sub-domain meets clustering requirements; and

a first clustering sub-module, which causes the device to cluster the sub-domain of the domain name of the phishing website into the cluster information library when it is determined that the count result of the sub-domain meets clustering requirements.

8. The device according to claim 7, wherein the clustering module comprises:

a second addition sub-module, which causes the device to obtain a count result of the domain name of the phishing website after incrementing by 1, when it is determined that the domain name type does not pertain to the SLD name;

a third determining sub-module, which causes the device to determine whether the count result of the domain name of the phishing website satisfies clustering requirements; and

a second clustering sub-module, which causes the device to cluster the domain name of the phishing website into the cluster information library when it is determined that the count result of the domain name of the phishing website satisfies clustering requirements.

9. The device according to claim 7, wherein the clustering requirements comprises one of:

the count result of the sub-domain exceeds a first preset threshold value within a first preset time; or

a ratio of the count result of the sub-domain to websites of an entire domain or to websites of the sub-domain exceeds a first preset ratio within the first preset time.

10. The device according to claim 8, wherein the clustering requirements comprises one of:

the count result of the domain name of the phishing website exceeds a second preset threshold value within a second preset time; or

a ratio of the domain name of the phishing website to websites of an entire domain or to websites of the sub-domain exceeds a second preset ratio within the second preset time.

11. The device according to claim 6, wherein the device operates as one or more of: a smart

phone, a personal computer, a laptop computer, a notebook computing device, a tablet computing device, a server or a terminal with a network connection.

12. A non-transitory computer-readable storage medium, wherein the computer readable storage medium stores a program which comprises codes or instructions to cause a processor circuitry to execute operations for clustering webpages, the operations comprising:

receiving a phishing website;

obtaining a domain name of the phishing website;

obtaining from a preset domain name table, a domain name type corresponding to the domain name of the phishing website ; and

clustering phishing webpages according to the domain name type.

13. The non-transitory computer-readable storage medium according to claim 12, wherein the clustering of phishing webpages according to the domain name type comprises:

determining whether the domain name type pertains to a secondary level domain (SLD) name, and if yes, obtaining a sub-domain of the domain name of the phishing website;

when a preset cluster information library does not include the sub-domain, obtaining a count result of the sub-domain after incrementing by 1; and

determining whether the count result of the sub-domain meets clustering requirements, and if yes, clustering the sub-domain of the domain name of the phishing website into the cluster information library.

14. The non-transitory computer-readable storage medium according to claim 12, wherein the method further comprises:

when the domain name type does not pertain to the SLD name, obtaining a count result of the domain name of the phishing website after incrementing by 1; and

determining whether the count result of the domain name of the phishing website satisfies clustering requirements, and if yes, clustering the domain name of the phishing website into the cluster information library.

15. The non-transitory computer-readable storage medium according to claim 13, wherein the clustering requirements comprises one of:

the count result of the sub-domain exceeds a first preset threshold value within a first preset time; or

a ratio of the count result of the sub-domain to websites of an entire domain or to websites of the sub-domain exceeds a first preset ratio within the first preset time.

16. The non-transitory computer-readable storage medium according to claim 14, wherein the clustering requirements comprises one of:

the count result of the domain name of the phishing website exceeds a second preset threshold value within a second preset time; or

a ratio of the domain name of the phishing website to websites of an entire domain or to websites of the sub-domain exceeds a second preset ratio within the second preset time.

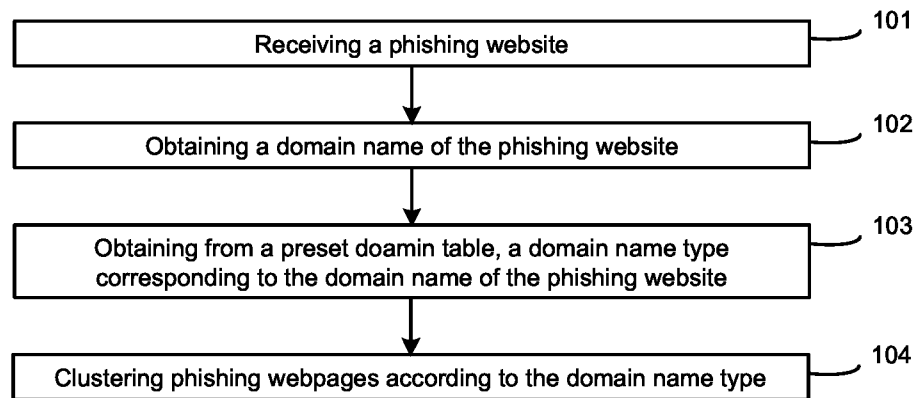


Figure 1

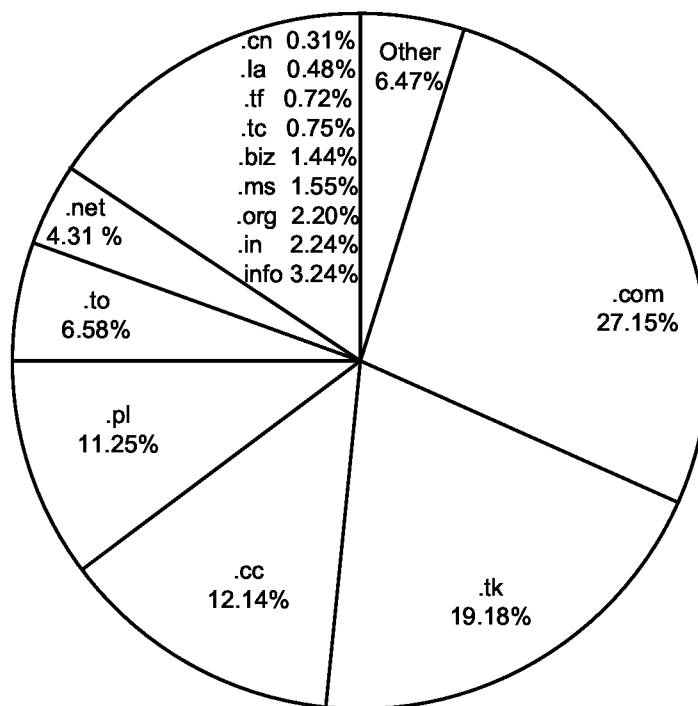


Figure 2

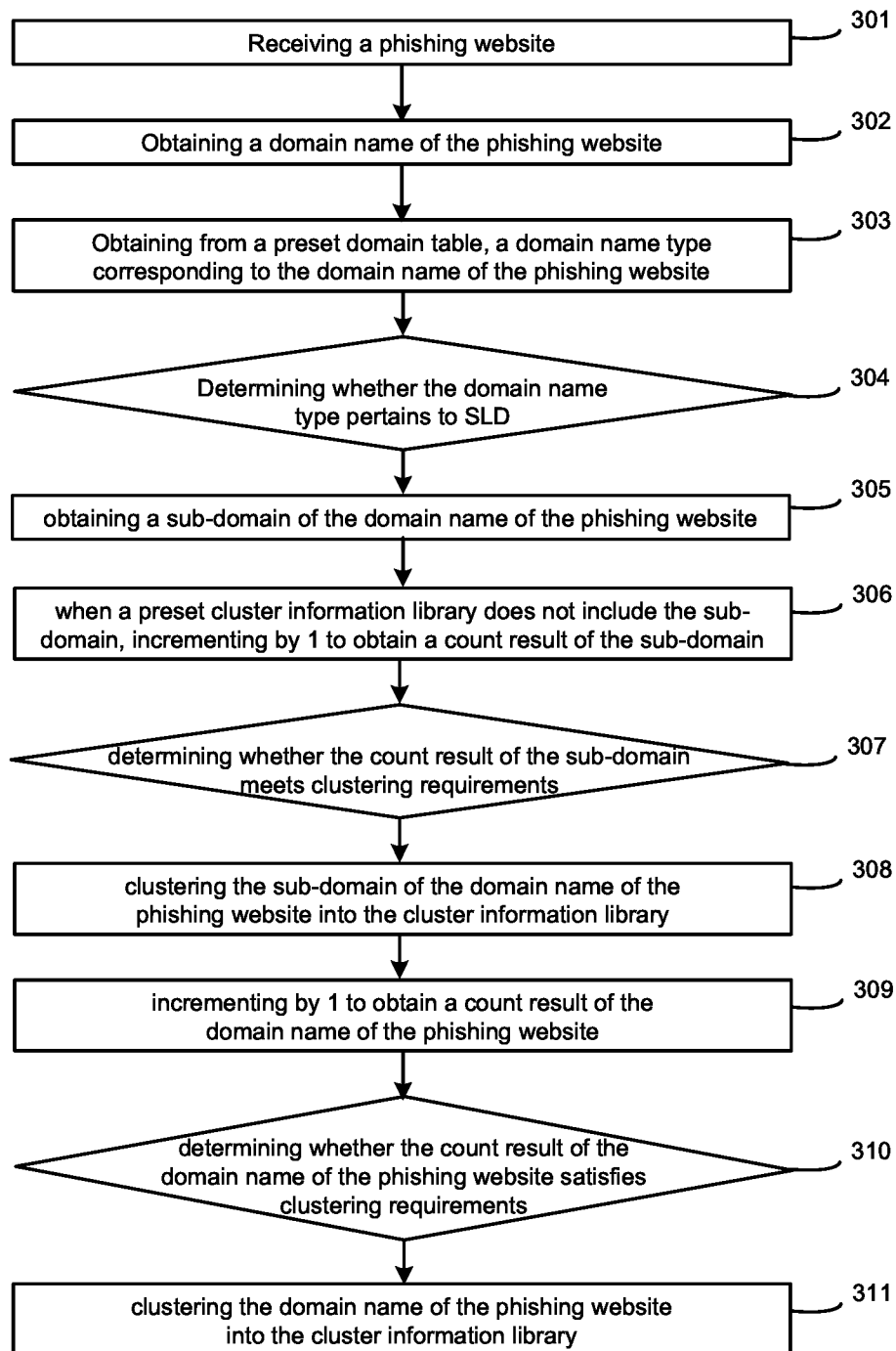


Figure 3

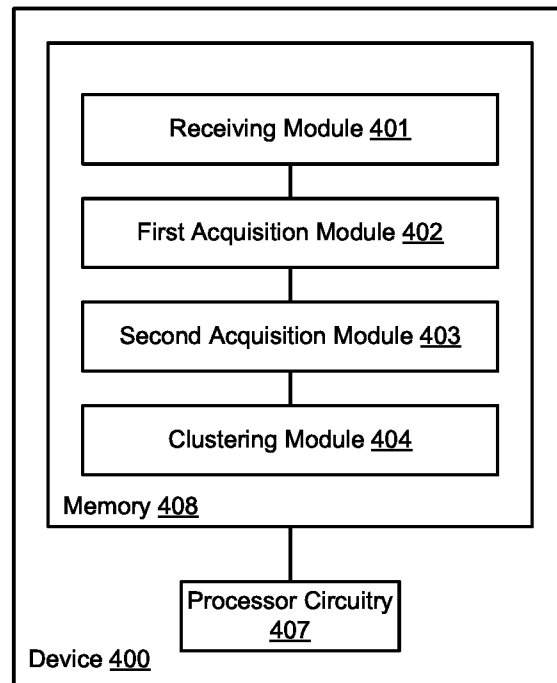


Figure 4

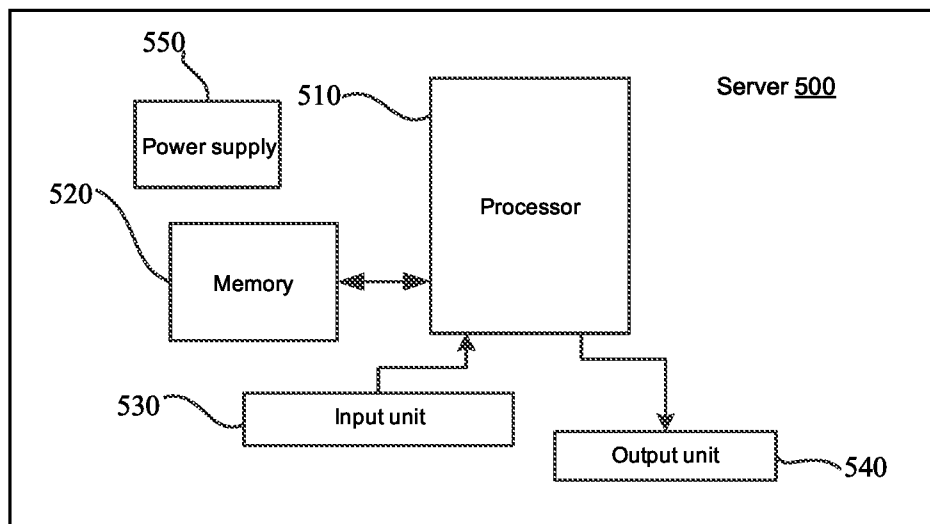


Figure 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2014/083261

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)

G06F

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)

CNKI, CNPAT, WPI, EPODOC:sub, second+, level, domain, web+, fraud+, phishing, counterfeit, malignant, black

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 103399912 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.) 20 November 2013 (2013-11-20) claims 1-7	1-16
X	CN 102938769 A (STATE COMPUTER NETWORK&INFORMATION SAF. ET AL.) 20 February 2013 (2013-02-20) description, paragraphs [0019]-[0026], claims 1-6 and figure 1	1, 6, 11, 12
A	CN 102571404 A (BEIJING D-MEDIA COMMUNICATION TECHNOLOGY) 11 July 2012 (2012-07-11) the whole document	1-16
A	US 2006123464 A1 (MICROSOFT CORPORATION) 08 June 2006 (2006-06-08) the whole document	1-16
A	US 7698442 B1 (VOLTAGE SECURITY, INC.) 13 April 2010 (2010-04-13) the whole document	1-16



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

13 October 2014

Date of mailing of the international search report

03 November 2014

Name and mailing address of the ISA/CN

STATE INTELLECTUAL PROPERTY OFFICE OF THE
P.R.CHINA(ISA/CN)
6,Xitucheng Rd., Jimen Bridge, Haidian District, Beijing
100088 China

Authorized officer

MU,Ying

Facsimile No. (86-10)62019451

Telephone No. (86-10)62413989

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2014/083261

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	103399912	A	20 November 2013	Non	e		
CN	102938769	A	20 February 2013	Non	e		
CN	102571404	A	11 July 2012	Non	e		
US	2006123464	A1	08 June 2006	US	7634810	B2	15 December 2009
				US	2007033639	A1	08 February 2007
				US	2006123478	A1	08 June 2006
				US	2007039038	A1	15 February 2007
				US	8291065	B2	16 October 2012
US	7698442	B1	13 April 2010	Non	e		