

METHOD AND SYSTEM FOR IDENTIFYING CLUSTERS OF AFFILIATED WEB RESOURCES

FIELD

[0001] The present technology generally relates to the field of information security; and more specifically, to methods and systems for identifying clusters of affiliated web resources.

BACKGROUND

[0002] Fraud is the most common and easiest way of committing cybercrimes. For example, according to the statistics provided by Group-IB, losses from a single phishing attack may amount to around \$20,000. In addition to substantial financial loss, damage to reputation should be considered.

[0003] There are two types of phishing targets: individual users and companies. Typically, in the former scenario, online fraudsters tend to operate alone because performing such an attack can be uncomplicated and may not require any specific expertise. This type of attacks becomes even more common due to the "phishing kits" available for purchase in the darknet (that is, illegally).

[0004] The attacks of the latter type, that is, those targeting companies and brands, may be performed in a different way. Typically, for a successful phishing attack against a targeted web resource, such as that of a bank, for example, the online fraudsters may have specific knowledge of programming and social engineering to scout for the initial target. As a result, the online fraudsters may organize in groups and associations.

[0005] According to the estimates of Group-IB, in Russia alone, there are approximately 15 online fraudster associations specializing in phishing, and this number is growing. Each of these associations can simultaneously own several hundred phishing pages targeted at different valid web resource – such as those representing companies or brands associated therewith.

[0006] Generally, a given phishing page targeting a given valid web resource may be created using a common pattern. To that end, the given phishing web resource may exist only for a predetermined period and may further be replaced by a new phishing web resource corresponding to the same common pattern and still targeting the given valid web resource, but with a different hosting provider or associated with a different email, for example. As soon as each new phishing page appears, urgent action is required to prevent the financial losses of the given valid web resource.

[0007] Certain prior art approaches have been proposed to address the above-identified technical problem.

[0008] **Russian Patent No.: 2681699-C1** issued on March 12, 2019, assigned to Trust LLC, and entitled “*Method and Server for Searching Related Network Resources*” discloses a method including scanning the network in order to search for network resources; in the network scanning stage, at least the first network resource and at least the second network resource are found; retrieving information about the found at least the first network resource and at least the second network resource including at least one parameter of the first network resource and at least one parameter of the second network resource; in response to the fact that at least one parameter of the first network resource coincides with at least one parameter of the second network resource, building a connection between the first network resource and the second network resource.

[0009] **PCT Application Publication No.: 2019/010182-A1** published on January 10, 2019, assigned to CLEVELAND *et al.*, and entitled “*Method and System for Detecting Phishing*” discloses a method of detecting a phishing event comprises acquiring an image of visual content rendered in association with a source, and determining that the visual content includes a password prompt. The method comprises performing an object detection, using an object detection convolutional network, on a brand logo in the visual content, to detect one or more targeted brands. Spatial analysis of the visual content may be performed to identify one or more solicitations of personally identifiable information. The method further comprises determining, based on the object detection and the spatial analysis, that at least a portion of the visual content resembles content of a candidate brand, and comparing the domain of the source with one or more authorized domains of the candidate brand. A phishing event is declared when the comparing indicates that the domain of the source is not one of the authorized domains of the candidate brand.

[0010] **United States Application Publication No.: 2016/0055490-A1** published on February 25, 2016, assigned to BRANDSHIELD LTD, and entitled “*Device, System, and Method Of Protecting Brand Names and Domain Names*” discloses a A computerized method of protecting a brand name of a brand owner, includes: (a) crawling a global communication network to identify and collect data about web-sites that possibly abuse the brand name; (b) for each web-site that possibly abuses the brand name, analyzing whether or not the web-site abuses the brand name by analyzing at least one of: (i) content of the web-site; and (ii) data about an owner of the web-site. The method further includes: for each web-site that possibly abuses the brand name, (A) generating an investment score indicating an estimated level of investment that was invested in development of the web-

site; and (B) generating a damage score indicating a level of damage that the web-site is estimated to produce to the brand name.

SUMMARY

[0011] It is an object of the present technology to ameliorate at least some of the inconveniences present in the prior art.

[0012] Developers of the present technology have realized that the prior art approaches may be somewhat cumbersome as they may have to consider too many web resources for generating clusters thereof, which eventually may affect efficiency and effectiveness of these approaches. In addition, the aforementioned prior art approaches appear to be directed to generating the clusters of web resources based on separate specific parameters, such as a DNS address associated with the given phishing web resource, which may not allow effectively responding to phishing attacks generated by web resources affiliated to the given phishing web resource.

[0013] Therefore, the developers have appreciated that the clusters of the web resources may be generated based on specific attributes associated therewith. In this regard, each cluster of web resources may include, for example, those web resources that have been created by a same owner/group, have same structural elements placed on similar positions within the associated web pages, associated with similar regional settings, and the like. By so doing, the clusters of web resources may be used to detect new web resources affiliated with one of the web resources therefrom more effectively and in a timely manner. This, in turn, may allow developing a more efficient strategy for identifying and preventing phishing attacks.

[0014] Thus, according to a first broad aspect of the present technology, there is provided a method of determining an affiliation of a web resource to a plurality of clusters of web resources. The method is executable by a computing device having at least a processing unit and a memory. The memory stores computer executable instructions, which, when executed, cause the processing unit to execute: at a training stage: scanning a communication network to detect a plurality of web resources; retrieving information associated with the plurality of web resources; based on the information, for each one of the plurality of web resources, generating a respective pattern; based on the respective pattern, grouping the plurality of web resources into the plurality of clusters, the grouping being based on a pattern affiliation between the respective patterns; storing an indication of the plurality of clusters of web resources in the memory; at a run-time stage: receiving an indication of an URL to a given web resource; retrieving the information about the given web resource; generating a new pattern of the given web resource; analyzing the pattern affiliation of

the new pattern with the patterns associated with the plurality of clusters of the web resources; calculating an affiliation ratio with at a specific one of the plurality of clusters; in response to the affiliation ratio with the specific one of the plurality of clusters exceeding a predetermined threshold value, associating the given web resource with the specific one of the plurality of clusters; updating the plurality of clusters stored in the database based on the given web resource.

[0015] In some implementations of the method, the respective pattern comprises a matrix of distinctive attributes associated with a respective one of the plurality of web resources identified via analyzing the information associated with the plurality of web resources.

[0016] In some implementations of the method, the distinctive attributes associated with the respective one of the plurality of web resources comprise at least one of: a size of at least one page block within at least one page associated with the respective one of the plurality of web resources; a position of the at least one page block within the at least one page; a title associated with the at least one page block; region parameters associated with the at least one page; at least one target domain associated with the at least one page, including at least one of: links placed within the at least one page and transition types associated therewith, the transition types further including at least one of: a direct transition type and a redirect transition type; contact details; path patterns to respective structural elements of the at least one page and styles associated therewith; and names of the structural elements.

[0017] In some implementations of the method, the information associated with the plurality of web resources includes, for a given web resource, at least one of: an URL of the given web resource; an HTML code associated with the given web resource; a screenshot of at least one web page associated with the given web resource page and a hash value thereof; a date of detecting the given web resource; and domain registration data associated with the given web resource, including: a registration date, a registrar, an owner name and contact details; an IP address; an NS server; a hosting provider; a last activity date.

[0018] In some implementations of the method, the matrix of distinctive attributes contains at least one distinctive attribute associated with the respective one of the plurality of web resources.

[0019] In some implementations of the method, each of the distinctive attributes has been selected based on a pre-determined parameter R , the pre-determined parameter being indicative of a maximum threshold value for a degree of distinctiveness of a given distinctive attribute for selection thereof for the respective pattern.

[0020] In some implementations of the method, the degree of distinctiveness of the given distinctive attribute is determined based on the following inequality:

$$p_i < R,$$

where p is a number of web resource within the plurality of web resources associated with the given distinctive attribute.

[0021] In some implementations of the method, the calculating the affiliation ratio further comprises calculating a number of distinctive attributes associated with the given web resource that are similar to those associated with the specific one of the plurality of clusters.

[0022] In some implementations of the method the analyzing the pattern affiliation of the new pattern with the patterns associated with the plurality clusters further comprises applying a cross-correlation technique.

[0023] According to a second broad aspect of the present technology, there is provided a system for determining an affiliation of a web resource to a plurality of clusters of web resources. The system comprises a computing device, the computing device further comprising: a processor; a non-transitory computer-readable medium comprising instructions. The processor, upon executing the instructions, is configured to: at a training stage: scan a communication network to detect a plurality of web resources; retrieve information associated with the plurality of web resources; based on the information, for each one of the plurality of web resources, generate a respective pattern; based on the respective pattern, group the plurality of web resources into the plurality of clusters, the grouping being based on a pattern affiliation between the respective patterns; store an indication of the plurality of clusters of web resources in the memory; at a run-time stage: receive an indication of an URL to a given web resource; retrieve the information about the given web resource; generate a new pattern of the given web resource; analyze the pattern affiliation of the new pattern with the patterns associated with the plurality of clusters of web resources; calculate an affiliation ratio with at a specific one of the plurality of clusters; in response to the affiliation ratio with the specific one of the plurality of clusters exceeding a predetermined threshold value, associate the given web resource with the specific one of the plurality of clusters; update the plurality of clusters stored in the database based on the given web resource.

BRIEF DESCRIPTION OF THE DRAWINGS

[0024] Non-limiting embodiments of the present technology are described herein with reference to the accompanying drawings; these drawings are only presented to explain the essence of the technique and are not limiting the scope of the present technology in any way, where:

[0025] Figure 1 depicts a schematic diagram of an example system suitable for implementation certain non-limiting embodiments of the present technology;

[0026] Figure 2 depicts a flowchart diagram of a method for generating clusters of web resources, in accordance with certain non-limiting embodiments of the present technology;

[0027] Figure 3 depicts a schematic diagram of an example pattern associated with a given web resource and generated in accordance with the method of Figure 2, in accordance with certain non-limiting embodiments of the present technology;

[0028] Figures 4A to 4B depict example web pages having been selected for pattern affiliation analysis in accordance with the method of Figure 2, in accordance with certain non-limiting embodiments of the present technology;

[0029] Figure 4C depicts an example common pattern having been generated for the web pages of Figures 4A and 4B in accordance with the method of Figure 2;

[0030] Figure 5 depicts a flowchart diagram of a method for determining an affiliation of a new resource to the clusters of web resources generated in accordance with the method of Figure 2, in accordance with certain non-limiting embodiments of the present technology;

[0031] Figure 6 shows a schematic diagram of an example computing environment of the system of Figure 1 configurable for execution of methods of Figure 2 and 5, in accordance with certain non-limiting embodiments of the present technology.

DETAILED DESCRIPTION

[0032] The following detailed description is provided to enable any one skilled in the art to implement and use the non-limiting embodiments of the present technology. Specific details are provided merely for descriptive purposes and to give insights into the present technology, and no was as a limitation. However, it would be apparent for the person skilled in the art that some of these specific details may not be necessary to implement certain non-limiting embodiments of the present technology. The descriptions of specific implementations are only provided as

representative examples. Various modifications of these embodiments may become apparent to the person skilled in the art; the general principles defined in this document may be applied to other non-limiting embodiments and implementations without departing from the scope of the present technology.

[0033] Certain non-limiting embodiments of the present technology are directed to systems and methods to identifying a plurality of cluster of web resources and determining affiliation of other web resources thereto.

[0034] With reference to Figure 1, there is depicted a schematic diagram of a system 100 used to determine the plurality clusters of web resources, according to certain non-limiting embodiments of the present technology

[0035] According to some non-limiting embodiments of the present technology, the system 100 comprises a data acquisition module 110 that may be communicatively coupled to a network 120.

[0036] In some non-limiting embodiments of the present technology, the network 120 is the Internet and/or an Intranet. Multiple embodiments of the network 120 may be envisioned and will become apparent to the person skilled in the art of the present technology. Further, how a communication link between the system 100 and the network 120 is implemented will depend, *inter alia*, on how the system 100 is implemented, and may include, but is not limited to, a wire-based communication link and a wireless communication link (such as a Wi-Fi communication network link, a 3G/4G communication network link, and the like).

[0037] Thus, in some non-limiting embodiments of the present technology, the data acquisition module 110 can scan the network 120 to receive respective URL links to a plurality of web resources and store them in an internal database (not depicted) communicatively coupled to the data acquisition module 110.

[0038] In alternative non-limiting embodiments of the present technology, the data acquisition module 110 may be configured to communicate with an external data storage (not depicted) to receive the respective URL links to the plurality of web resources for further analysis.

[0039] Further, having received the URL link to the given web resource, the data acquisition module may be configured to transmit it to a computing device (such as a computing device 600 depicted of Figure 6, not depicted in Figure 1). According to some non-limiting embodiments of the present technology, the computing device 600 can be configured to implement a method for classifying and filtering the content received via the network 120. The so processed data, by the

computing device 600, may further be transmitted back to the system 100. To that end, in some non-limiting embodiments of the present technology, the system 100 may further comprise an output module 130 communicatively coupled to the computing device 600.

[0040] In these embodiments, the output module 130 can be configured to output the analysis report associated with the URL link. This report may include data indicative of an affiliation score associated the given web resource which is located under the URL link received by the data acquisition module 110. How the affiliation score for the given web resource is calculated will be described immediately below.

[0041] With reference to Figure 2, there is provided a flowchart diagram of a method 200 for generating clusters of web resources, according to certain non-limiting embodiments of the present technology. The method 200 may be executed by the computing device 600.

STEP 210: SCANNING A COMMUNICATION NETWORK TO DETECT A PLURALITY OF WEB RESOURCES

[0042] The method 200 commences at step 210, where the computing device 600 is configured to scan the network 120 to identify the plurality web resources for further analysis. How the computing device 600 is configured to scan the network 120 is not limited, and in certain non-limiting embodiments of the present technology, may be performed using any known specific software for scanning and crawling communication networks.

[0043] In alternative non-limiting embodiments of the present technology, the web resources may be received, by the computing device 600, from a database of web resources, which may be, for example, internally hosted at the computing device 600 or communicatively coupled thereto via the network 120.

[0044] The method 200 hence advances to step 220.

STEP 220: RETRIEVING INFORMATION ASSOCIATED WITH THE PLURALITY OF WEB RESOURCES;

[0045] At step 220, according to the certain non-limiting embodiments of the present technology, once the URL link to the given web resource has been received, the information associated therewith may thus be retrieved.

[0046] In some non-limiting embodiments of the present technology, the computing device 600 may be configured to retrieve the information associated with the given web resource including at least one of:

- a URL of the given web resource;
- an HTML code of at least one page associated with the given web resource;
- a screenshot of the at least one page associated with the given web resource and a hash value thereof;
- a date of detecting the given web resource; and domain registration data associated with the given web resource, further including:
 - a registration date, a registrar,
 - an owner name and contact details;
 - an IP address;
 - an NS server;
 - a hosting provider;
 - a last activity date.

[0047] Further, in some non-limiting embodiments of the present technology, the computing device 600 may be configured to store the information associated with the given web resource in the database of the web resources.

[0048] According to some non-limiting embodiments of the present technology, the computing device 600 may further be configured to analyze the HTML code of at least one page associated with the given web resource to identify distinctive attributes associated with the given web resource.

[0049] In some non-limiting embodiments of the present technology, the computing device 600 may be configured to conduct such an analysis of the HTML code associated with given the web resource pages automatically, e.g., using a specialized software which can detect the distinctive attributes associated with the given web resource.

[0050] In additional non-limiting embodiments of the present technology, the analyzing, by the computing device 600, the HTML code associated with the given web resource may be performed to identify distinctive attributes of pages associated with the given web resource, which may further be used to analyze similarity of patterns thereof.

[0051] According to certain non-limiting embodiments of the present technology, the distinctive attributes identified via the analyzing the HTML code associated with the given web resource page may include at least one of:

- a size of at least one page block within at least one page associated with the respective one of the plurality of web resources;
- a position of the at least one page block within the at least one page;
- a title associated with the at least one page block;
- regional parameters associated with the at least one page;
- at least one target domain associated with the at least one page, including at least one of: links placed within the at least one page and transition types associated therewith, the transition types further including at least one of: a direct transition type and a redirect transition type;
- contact details – for example, contact details associated with a registrar of the at least one page associated with the given web resources, or those placed therewithin;
- path patterns to respective structural elements of the at least one page and styles associated therewith; and
- a name of each of the respective structural elements.

[0052] In some non-limiting embodiments of the present technology, each of the aforementioned distinctive attributes may be represented by a respective sequence of bytes in the HTML code of the at least page, or may contain a sequence of bytes indicative of one of a brand and a phishing target of the given web resource.

[0053] In some non-limiting embodiments of the present technology, the computing device 600 may be configured to store, in the database of web resources, a hash value of the at least one page associated with the given web resource. It should be expressly understood that, according to the

non-limiting embodiments of the present technology, the hash value may be obtained using any suitable hashing algorithm, such as a universal hash function, a non-cryptographic hash function, a keyed cryptographic hash function, and the like.

[0054] In the context of the present specification, a screenshot of a given web page denotes a graphic representation (such as a picture) of a content of the given web page. Such a graphic representation may be stored as a file in the database of the web resources associated with the computing device 600. Further, data indicative of the screenshot, along with an associated hash values, may be used to estimate an affiliation score associated therewith. The screenshots may be compared both bitwise and based on hash values of associated pages.

[0055] In some non-limiting embodiments of the present technology, the computing device 600 may further be configured to determine the domain registration data including at least the one of: a registration date, a registrar, an owner's name and contact details, as an example.

[0056] It should be noted that how the domain registration data may be acquired is not limited; and, in specific non-limiting embodiments of the present technology, to determine the domain registration data, the computing device 600 may be configured to have access, via the network 120, to a WHOIS domain lookup service. Typically, the WHOIS domain lookup service may be configured to provide, in response to a respective WHOIS query, detailed information about the domain of the given web resource, including, but not being limited to: a registration date and time, an expiry date and time, a current DNS sever of the domain, its statuses, and also information about the registrar that owns the domain name. If the domain name does not exist, the response to the respective WHOIS query may indicate that the domain name has not been found. In certain scenarios, if the domain zone does not provide the information about the owner of the domain name in the response to the respective WHOIS query, the WHOIS domain lookup service may be configured to request this information from the domain registrar.

[0057] The method 200 hence advances to step 230.

STEP 230: BASED ON THE INFORMATION, FOR EACH ONE OF THE PLURALITY OF WEB RESOURCES, GENERATING A RESPECTIVE PATTERN;

[0058] At step 230, having retrieved the necessary information associated with the given web resource, the computing device 600 may further be configured to generate a respective pattern for the given web resource.

[0059] In some non-limiting embodiments of the present technology, the respective pattern for the given web resources may comprise a matrix of the distinctive attributes associated with the given web resource.

[0060] With reference to Figure 3, there is depicted a schematic diagram of the a pattern 302 generated, by the computing device 600, for the given web resource, in accordance with certain non-limiting embodiments of the present technology.

[0061] As it can be appreciated, the pattern 302 has been generated, by the computing device 600, based on the information acquired during the analysis of the given web resource, thereby identifying the distinctive attributes associated therewith.

[0062] As mentioned earlier, in some non-limiting embodiments of the present technology, the information associated with the given web resource may include a position and size of at least one page block within at least page associated with the given web resource.

[0063] Below is an example excerpt of the HTML code of the at least one page associated with the given web resource representative of a position of "block_1" placed therewithin:

```
#block_1 {  
  
top: 82px; /* Distance from the top edge */  
  
left: 170px; /* Distance from the left edge */  
  
right: 73px; /* Distance from the right edge */  
  
bottom: 0/* Distance from the bottom */  
  
height: 80px; /* Height*/  
  
width: 150px; /*Width*/  
  
}
```

[0064] According to this example, the block_1 is positioned, on the at least one page of the given web resource, at 82 pixels below the top edge, at 170 pixels from the left edge, at 73 pixels from the right edge and at 0 pixels from the bottom; and it is 150 pixels wide and 80 pixels high.

[0065] Further, in some non-limiting embodiments of the present technology, the information associated with the given web resource may include the registration date and the domain registrar associated the domain associated with the given web resource.

[0066] For example, in response to the respective WHOIS query, as described above in respect of execution of step 220, the WHOIS domain lookup service may be configured to provide the domain registration data as presented in Table 1 below. As it can be appreciated from Table 1, GROUP-IB.RU domain was registered on June 14, 2007 and its registrar is REGRU-RU

Table 1

Domain	GROUP-IB.RU
DNS Server	<u>dora.ns.cloudflare.com.</u>
DNS Server	<u>tom.ns.cloudflare.com.</u>
Registrar	REGRU-RU
Registration Date	2007-06-14T20:00:00Z
Registration Expiry Date	2020-06-14T21:00:00Z

[0067] As previously mentioned, in other non-limiting embodiments of the present technology, the computing device 600 may further be configured to retrieve links placed within the at least one page and transition types associated therewith, the transition types further including at least one of: a direct transition type and a redirect transition type; and contact details associated with the at least one web page.

[0068] Broadly speaking, in some non-limiting embodiment of the present technology, the respective pattern may be generated based on any combination of distinctive attributes identified at step 220. Therefore, in some non-limiting embodiments of the present technology, the pattern 302 associated with the given web resource may include at least one distinctive attribute to be used for further analysis.

[0069] With reference to Figures 4A to 4B, there is schematically depicted pattern affiliation between two affiliated web resources respectively represented by a first page 402 (depicted in Figure 4A) and a second page 404 (depicted in Figure 4B), in accordance with certain non-limiting embodiments of the present technology.

[0070] Further, with reference to Figure 4C, there is depicted a common pattern 406 having been generated, by the computing device 600, for the first page 402 and the second page 404 depicted in Figures 4A and 4B, respectively, based on pattern affiliation analysis thereof, in accordance with certain non-limiting embodiments of the present technology.

[0071] In accordance with certain non-limiting embodiments of the present technology, each of the first page 402 and the second page 404 can have similar respective patterns (not separately labelled), however different within a certain confidence interval. In this regard, based on the pattern affiliation analysis described herein below in respect of step 240 of the method 200, the computing device 600 may be configured to generate the common pattern 406 for both the first page 402 and the second page 404, thereby determining affiliation therebetween. In some non-limiting embodiments of the present technology, the common pattern 406 may further be modified and updated, if necessary, such as upon identification new distinctive attributes associated with the first page and the second page, as an example, as will be described below.

[0072] Thus, the method 200 further proceeds to step 240.

STEP 240: BASED ON THE RESPECTIVE PATTERN, GROUPING THE PLURALITY OF WEB RESOURCES INTO THE PLURALITY OF CLUSTERS, THE GROUPING BEING BASED ON A PATTERN AFFILIATION BETWEEN THE RESPECTIVE PATTERNS;

[0073] At step 240, the plurality of web resources, received at step 220 of the method 200, are further grouped (or otherwise clustered) based on a predetermined parameter indicative of affiliation of their respective patterns, thereby generating a plurality of clusters of the web resources.

[0074] To that end, the computing device 600 may be configured to determine, for the given web resources, a set of distinctive attributes to be included in the respective pattern. According to certain non-limiting embodiments of the present technology, the computing device 600 may be configured to determine the set of distinctive attributes based on their associated degrees of distinctiveness for the given resource.

[0075] In some non-limiting embodiments of the present technology, to determine a degree of distinctiveness, for each of the distinctive attributes associated with the given web resource, a predetermined parameter R may be used. In these embodiments, the predetermined parameter R may be determined by way of experiment using test samples to identify distinctive attributes for predetermined clusters of web resources as precisely as possible. In some non-limiting embodiments of the present technology, the predetermined parameter R may have been determined by a machine-learning algorithm trained based on the test samples.

[0076] According to some non-limiting embodiments of the present technology, the predetermined parameter R may be selected to exceed an estimated maximum size of a given group of web resources. For example, a value of the predetermined parameter R may be calculated according to the following formula:

$$R=0.7 \cdot V, \quad (1)$$

where V is a volume of the database of web resources associated with the computing device 600.

[0077] It should further be noted that other coefficients, such as 0.3, 0.5, 0.8, or 1.2, for example, may also be used for determining the value of the predetermined parameter R depending on a particular implementation of the present technology.

[0078] Further, for each distinctive attribute, the computing device 600 may be configured to determine a respective value of p corresponding to a number of web resources in the database of web resources associated with this distinctive attribute.

[0079] Additionally, the respective value of p may be selected in such a way that it does not exceed the value of the predetermined parameter R, which corresponds to the estimated maximum size of the given group of web resources, that is, the respective value of p may satisfy the following inequality:

$$p_i < R. \quad (2)$$

[0080] Also, those distinctive attributes associated with respective values of p that are less than the value of R may be considered (sufficiently) distinctive for the given resource and thus will be selected in the set of distinctive attributes for generating the respective pattern.

[0081] By doing so, in some non-limiting embodiments of the present technology, the respective pattern may comprise a matrix of distinctive attributes associated with the given web resource. By

doing so, the computing device 600 may be configured to generate a respective matrix for each one of the plurality of web resources identified at step 220 of the method 200 as described above.

[0082] Further, in some non-limiting embodiments of the present technology, the computing device 600 may be configured to analyze the respective matrices associated with the plurality of web resources amongst each other. To that end, in some non-limiting embodiments of the present technology, the analyzing may comprise comparing the respective matrices amongst each other by applying a cross-correlation technique.

[0083] In the context of the present specification, the term “cross-correlation technique” relates to the field of probability and statistics and refers to a method of determining similarities between entries of two sets of data, such as the distinctive attributes of the respective matrices associated with two given web resources – for example, by constructing a cross-correlation function describing similarities between the two sets of data.

[0084] Thus, according to certain non-limiting embodiments of the present technology, the computing device 600 may be configured to determine a proportion of matched distinctive attributes between the two given web resources, thereby determining an affiliation ratio t therebetween.

[0085] Generally, the respective matrices may be of different sizes. In this situation, the affiliation ratio may be calculated as a ratio between the number of matched distinctive attributes associated with the given resource to a number of distinctive attributes associated with one of the plurality of web resources having a smallest pattern.

[0086] In some non-limiting embodiments of the present technology, the computing device 600 may be configured to apply the cross-correlation technique simultaneously for at least some of the plurality of web resources. In other non-limiting embodiments of the present technology, the computing device 600 may be configured to apply the cross-correlation technique stepwise to each pair of the plurality of web resources.

[0087] Thus, by doing so, according to certain non-limiting embodiments of the present technology, the computing device 600 may be configured to generate the plurality of clusters for grouping therein the plurality of web resources.

[0088] The method 200 hence advances to step 250.

STEP 250: STORING AN INDICATION OF THE PLURALITY OF CLUSTERS OF WEB RESOURCES IN THE MEMORY

[0089] Finally, according to certain non-limiting embodiments of the present technology, at step 250, the computing device 600 may be configured to store the plurality of clusters of the web resources in one of the database of web resources associated with the computing device 600 and the internal database of the system 100.

[0090] The method 200 hence terminates.

[0091] According to certain non-limiting embodiments of the present technology, having generated the plurality of clusters for the plurality of web resources, the computing device 600 may further be configured to determine affiliation of a new web resource to the so identified plurality of clusters. To that end, the computing device 600 may be configured to execute a method 300, a flowchart diagram of which is depicted in Figure 5.

STEP 310: RECEIVING AN INDICATION OF AN URL TO A GIVEN WEB RESOURCE

[0092] At step 310, the computing device may be configured to receive a new URL associated with a new web resource to be analyzed for affiliation with the plurality of clusters of the web resources. The computing device 600 may be configured to receive the new URL as described above in respect of implementation of step 210 of the method 200.

STEP 320: RETRIEVING THE INFORMATION ABOUT THE GIVEN WEB RESOURCE

[0093] Further, at step 320, information about the new web resource is retrieved. Based on this information, the computing device 600 may be configured to identify distinctive attributes associated with the new web resource in a fashion similar to that described above with respect to step 220 of the method 200.

STEP 330: GENERATING A NEW PATTERN OF THE GIVEN WEB RESOURCE

[0094] Based on the received information, at step 330, the computing device 600 may further be configured to generate a new pattern associated with the new web resource located under the new URL link received at step 310 of the method 300.

STEP 340: ANALYZING THE PATTERN AFFILIATION OF THE NEW PATTERN WITH THE PATTERNS ASSOCIATED WITH THE PLURALITY OF CLUSTERS OF THE WEB RESOURCES

[0095] Thus, at step 340, the computing device 600 may be configured to analyze the new pattern against each one of the respective patterns associated with the plurality of web resources stored in the database of web resources associated with the computing device 600 and having been identified via execution of the method 200.

STEP 350: CALCULATING AN AFFILIATION RATIO WITH AT A SPECIFIC ONE OF THE PLURALITY OF CLUSTERS

[0096] At step 350, the computing device 600 may be configured to calculate a respective affiliation ratio between the new pattern and those associated with the plurality of clusters.

STEP 360: IN RESPONSE TO THE AFFILIATION RATIO WITH THE SPECIFIC ONE OF THE PLURALITY OF CLUSTERS EXCEEDING A PREDETERMINED THRESHOLD VALUE, ASSOCIATING THE GIVEN WEB RESOURCE WITH THE SPECIFIC ONE OF THE PLURALITY OF CLUSTERS

[0097] Further, at step 360, in response to a given affiliation ration, between the new pattern and that of a specific one of the plurality of clusters exceeds a predetermined threshold value, the computing device 600 may be configured to determine the new pattern and the new web resource associated therewith as being affiliated with the specific one of the plurality of clusters. Conversely, if the respective affiliation ratio does not exceed the predetermined threshold value, the new pattern may thus be identified as not affiliated with the specific one of the plurality of clusters.

STEP 370: UPDATING THE PLURALITY OF CLUSTERS STORED IN THE DATABASE BASED ON THE GIVEN WEB RESOURCE

[0098] Finally, at step 370, the computing device 600 may be configured to update the specific one of the plurality of clusters and the respective pattern associated therewith, for example, based on the distinctive attributes associated with the new web resource.

[0099] In additional non-limiting embodiments of the present technology, the computing device 600 may be configured to generate a detailed analysis report including data indicative of the affiliation of the new web resource relative to each of the plurality of clusters.

[0100] The method 300 hence terminates.

[0101] With reference to Figure 6, there is depicted an example functional diagram of the computing device 600 configurable to implement certain non-limiting embodiments of the present technology. computing device 600

[0102] In some non-limiting embodiments of the present technology, the computing device 600 may include: one or more central processing units (CPUs) 601, at least one non-transitory computer-readable memory 602, a storage 603, input/output interfaces 604, input/output peripherals 605, connectivity hardware 606.

[0103] According to some non-limiting embodiments of the present technology, the one or more CPUs 601 may be configured to execute specific program instructions the computations as required for the computing device 600 to function properly or to ensure the functioning of one or more of its components. The one or more CPUs 601 may further be configured to execute specific machine-readable instructions stored in the at least one non-transitory computer-readable memory 602, for example, those causing the computing device to execute one of the method 200 and the method 300.

[0104] The at least one non-transitory computer-readable memory 602 may be implemented as RAM and contains the necessary program logic to provide the requisite functionality.

[0105] The storage 603 may be implemented as at least one of an HDD drive, an SSD drive, a RAID array, a network storage, a flash memory, an optical drive (such as CD, DVD, MD, Blu-ray), etc. The storage 603 may be configured for long-term storage of various data, e.g., the aforementioned documents with user data sets, databases with the time intervals measured for each user, user IDs, etc.

[0106] The interfaces input/output 604 may comprise various interfaces, such as at least one of USB, RS232, RJ45, LPT, COM, HDMI, PS/2, Lightning, FireWire, etc.

[0107] The input/output peripherals 605 may include at least one of a keyboard, joystick, (touchscreen) display, projector, touchpad, mouse, trackball, stylus, speakers, microphone, and the like. . A communication link between the each one of the input/output peripherals 605 can be wired (for example, connecting the keyboard via a PS/2 or USB port on the chassis of the desktop PC) or wireless (for example, via a wireless link, e.g., radio link, to the base station which is directly connected to the PC, e.g., to a USB port).

[0108] The connectivity hardware 606 may be selected based on a particular implementation of the network 120, and may comprise at least one of: an Ethernet card, a WLAN/Wi-Fi adapter, a

Bluetooth adapter, a BLE adapter, an NFC adapter, an IrDa, a RFID adapter, a GSM modem, and the like. As such, the connectivity hardware 606 may be configured for wired and wireless data transmission, via one of WAN, PAN, LAN, Intranet, Internet, WLAN, WMAN, or GSM networks.

[0109] These and other components of the computing device 600 may be linked together using a common data bus 610.

[0110] It should be expressly understood that not all technical effects mentioned herein need to be enjoyed in each and every embodiment of the present technology.

[0111] Modifications and improvements to the above-described implementations of the present technology may become apparent to those skilled in the art. The foregoing description is intended to be exemplary rather than limiting. The scope of the present technology is therefore intended to be limited solely by the scope of the appended claims.

CLAIMS

1. A method of determining an affiliation of a web resource to a plurality of clusters of web resources, the method executable by a computing device having at least a processing unit and a memory, the memory storing computer executable instructions, which, when executed, cause the processing unit to execute:

at a training stage:

scanning a communication network to detect a plurality of web resources;

retrieving information associated with the plurality of web resources;

based on the information, for each one of the plurality of web resources, generating a respective pattern;

based on the respective pattern, grouping the plurality of web resources into a plurality of clusters, the grouping being based on a pattern affiliation between the respective patterns;

storing an indication of the plurality of clusters of web resources in the memory;

at a run-time stage:

receiving an indication of an URL to a given web resource;

retrieving the information about the given web resource;

generating a new pattern of the given web resource;

analyzing the pattern affiliation of the new pattern with the patterns associated with the plurality of clusters of the web resources;

calculating an affiliation ratio with a specific one of the plurality of clusters;

in response to the affiliation ratio with the specific one of the plurality of clusters exceeding a predetermined threshold value, associating the given web resource with the specific one of the plurality of clusters;

updating the plurality of clusters stored in the database based on the given web resource.

2. The method of claim 1, wherein the respective pattern comprises a matrix of distinctive attributes associated with a respective one of the plurality of web resources identified via analyzing the information associated with the plurality of web resources.

3. The method of claim 2, wherein the distinctive attributes associated with the respective one of the plurality of web resources comprise at least one of:

a size of at least one page block within at least one page associated with the respective one of the plurality of web resources;

a position of the at least one page block within the at least one page;

a title associated with the at least one page block;

region parameters associated with the at least one page;

at least one target domain associated with the at least one page, including at least one of: links placed within the at least one page and transition types associated therewith, the transition types further including at least one of: a direct transition type and a redirect transition type;

contact details;

path patterns to respective structural elements of the at least one page and styles associated therewith; and

names of the structural elements.

4. The method of claim 2, wherein the information associated with the plurality of web resources includes, for a given web resource, at least one of: an URL of the given web resource; an HTML code associated with the given web resource; a screenshot of at least one web page associated with the given web resource page and a hash value thereof; a date of detecting the given web resource; and domain registration data associated with the given web resource, including: a registration date, a registrar, an owner name and contact details; an IP address; an NS server; a hosting provider; a last activity date.

5. The method of claim 2, wherein the matrix of distinctive attributes contains at least one distinctive attribute associated with the respective one of the plurality of web resources.

6. The method of claim 2, wherein each of the distinctive attributes has been selected based on a pre-determined parameter R , the pre-determined parameter being indicative of a maximum threshold value for a degree of distinctiveness of a given distinctive attribute for selection thereof for the respective pattern.

7. The method of claim 6, wherein the degree of distinctiveness of the given distinctive attribute is determined based on the following inequality:

$$p_i < R,$$

where p is a number of web resource within the plurality of web resources associated with the given distinctive attribute.

8. The method of claim 1, wherein the calculating the affiliation ratio further comprises calculating a number of distinctive attributes associated with the given web resource that are similar to those associated with the specific one of the plurality of clusters.

9. The method of claim 1, wherein the analyzing the pattern affiliation of the new pattern with the patterns associated with the plurality clusters further comprises applying a cross-correlation technique.

10. A system for determining an affiliation of a web resource to a plurality of clusters of web resources, the system comprising a computing device, the computing device further comprising:

a processor;

a non-transitory computer-readable medium comprising instructions;

the processor, upon executing the instructions, being configured to:

at a training stage:

scan a communication network to detect a plurality of web resources;

retrieve information associated with the plurality of web resources;

based on the information, for each one of the plurality of web resources, generate a respective pattern;

based on the respective pattern, group the plurality of web resources into the plurality of clusters, the grouping being based on a pattern affiliation between the respective patterns;

store an indication of the plurality of clusters of web resources in the memory;

at a run-time stage:

receive an indication of an URL to a given web resource;

retrieve the information about the given web resource;

generate a new pattern of the given web resource;

analyze the pattern affiliation of the new pattern with the patterns associated with the plurality of clusters of web resources;

calculate an affiliation ratio with a specific one of the plurality of clusters;

in response to the affiliation ratio with the specific one of the plurality of clusters exceeding a predetermined threshold value, associate the given web resource with the specific one of the plurality of clusters;

update the plurality of clusters stored in the database based on the given web resource.