CLAIMS

1. A method for clustering phishing web resources based on visual content image, the method is executed on a computer device comprising at least a processor and a memory, the method comprises:

receiving references to multiple phishing web resources;

retrieving at least one image of the visual content of each phishing web resource from the multiple phishing web resources;

processing the each image associated with one phishing web resource from the multiple phishing web resources, including

identifying contours of elements in each image; and

filtering the identified contours in each image by removing all but one instance of the identical contours;

if a similarity value exceeds a threshold value: based on pairwise comparison of the identified contours and one or more contours of an existing cluster for a specific phishing set, combining in a database the phishing web resource associated with the identified contours and the existing cluster, otherwise, creating a new cluster in the database for the phishing web resource;

storing references to the phishing web resources in the database.

2. The method of claim 1, further comprising preliminary scanning a network for obtaining the stored references.

3. The method of claim 1, further comprising obtaining the stored references, at least partially, from an external source via a data transmission network.

4. The method of claim 1, further comprising obtaining the stored references from a preliminary-formed web resource database.

5. The method of claim 1, wherein the elements include: logos, letter outlines, individual words, pictures, and/or forms.

6. The method of claim 1, wherein the contours are identified using at least the following methods: Canny edge detector and/or contour tracing algorithm.

7. The method of claim 1, further comprising removing small-size, oblique and elongated contours when processing each image.

8. The method of claim 1, wherein filtering the identified contours in each image includes removing similar contours, and if any compared contours differ in height and/or width by more than two pixels, the contours are considered to be different.

9. The method of claim 1,

wherein filtering the identified contours in each image includes removing similar contours,

wherein, if the compared contours differ in height and/or width by two pixels or less, then the contours are overlapped, and the method further comprises calculating per-pixel difference of the images,

wherein, a similarity score for the resultant differential image is within a range from 0, where the images are the same, to 100, where the images are completely opposite,

wherein, the method further comprises calculating the similarity score (P) as the arithmetic mean of pixel values, that is, the sum of pixel values divided by the number of image pixels, and calculated for the values taken from all three RGB channels:

$$P = \frac{\sum_{i-1}^{N} Ri + \sum_{i-1}^{N} Gi + \sum_{i-1}^{N} Bi}{N},$$

where Ri, Gi, Bi – i-th pixel value in R, G and B channels, respectively, and N – total number of image pixels, and

wherein, the method further comprises removing all but one instance of the contours with a similarity score that does not exceed 7.

10. The method of claim 1, further comprising calculating invariant visual hash functions for each contour identified in the image.

11. The method of claim 1, further comprising displaying an image message if, when pairwise comparing the images, less than three similar contours are found,

and a new image is not added to the cluster, and if there are less than two images in the cluster, the cluster is deleted;

if the comparison reveals from three to five similar contours, then, in order to add an image to a specific cluster, all the cluster contours must match exactly;

if the comparison reveals from five to ten similar contours, then, in order to add an image to a specific cluster, at least half of the contours must match the cluster contours,

if the comparison reveals from ten to fifteen similar contours, then, in order to add an image to a specific cluster, at least a quarter of the contours must match the cluster contours,

if the comparison reveals fifteen or more similar contours, then, in order to add an image of the phishing web resource visual content to a specific cluster, at least one fifth of the contours must match the cluster contours.

12. The method of claim 1, wherein the phishing web resource is combined with the existing cluster with the largest number of contour matches.

13. The method of claim 1, wherein after clustering each of the images , the method further comprising deleting the clusters containing less than three images.

14. The method of claim 13, further comprising entering the images contained in the deleted clusters into a special list and are temporarily considered as garbage.

15. The method of claim 14, further comprising creating an adjacency matrix based on the number of matching contours for each image of the special list.

16. The method of claim 15, wherein, the number of matching contours in the adjacency matrix, is changed to 0, if the number of matching contours is less than a predetermined threshold, or to 1, if the number of matching contours is more than the threshold.

17. The method of claim 16, further comprising computing strongly connected graph components, in which there are more than two images, and creating a new cluster for each of the identified components.

18. The method of any preceding claim, further comprising transmitting the database to an external device.

19. The method of any preceding claim, further comprising storing the identified contours in the database.

20. A computing device for clustering phishing content, the device comprising a memory for storing machine-readable instructions and at least one computing processor configured to execute machine-readable instructions to ensure implementation of the method for clustering phishing content according to any of claims 1-19.