WebAA: Website Association Analysis via Multi-Resource Similarity Computation

Taiyao Zhang[1,2], Dongzheng Jia[3], Xingyu Fu[1,2,(✉)], Zhihao Zhang[1,2], and Qingyun Liu[1,2]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
{zhangtaiyao, fuxingyu, zhangzhihao, liuqingyun}@iie.ac.cn
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3] National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing, China
jdz@cert.org.cn

Abstract. The rapid proliferation of websites has posed significant challenges to cyberspace management. To effectively manage websites, we introduce the innovative concept of website association in this paper and propose a website association model, WebAA. Website association seeks to determine whether target websites belong to the same organization based on their features, playing a crucial role in cyberspace management. Unlike website identification, website association enables more accurate organizational alignment through fine-grained analysis of website features. Specifically, malicious website association helps investigators identify the organizations behind them, thereby addressing the root causes of the harm they cause. In simple terms, the proposed website association model in this paper consists of four modules. In the first module, we calculate the similarity score based on external resources by analyzing the dependency relationships of the target websites on these resources. The second module employs the BERT model to analyze the similarity of HTML texts between the target websites and generates a similarity score. The third module focuses on the website domain names, calculating their similarity using Levenshtein distance and obtaining the similarity score. In the fourth module, the final similarity score is obtained by weighting the three similarity scores, which is then used to determine whether the two websites belong to the same organization. Extensive experiments on two real-world datasets demonstrate that our model can efficiently associate thousands of website pairs within milliseconds with an accuracy exceeding 90%. Striving for applicability and replicability, we release ready-to-use raw data from our study.

Keywords: Web computing · Cyberspace governance · Website association · BERT.

1 Introduction

With the rapid development of the Internet, more and more organizations and institutions are expanding their influence by building their own websites. Accord-

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_4

ing to statistics from Worldwidewebsize [16], as of January 1, 2025, there are 3.76 billion indexable websites worldwide. The vast number of websites and the complex relationships between them pose challenges to cyberspace management, but this issue cannot be effectively addressed solely at the website level. Focusing management efforts on the organizational level behind the websites and conducting website management at that level is a feasible solution. Organizational-level website management enables management from an organizational perspective, revealing potential connections and interactive relationships, which leads to better management of cyberspace ecology.

Organizational-level website management method is highly valuable for fields like information security and data mining. Malicious attackers often create multiple associated websites to conceal their identities or expand their scope of influence [9]. For example, phishing websites, gambling websites, counterfeit platforms, and malicious advertising networks often operate in a distributed manner, using multiple associated websites to evade security detection and blocking [11]. The organization behind the website often uses a new domain name to quickly establish a replacement site when the current website is attacked [10]. Therefore, combating such malicious services solely from the perspective of individual websites is ineffective. Analyzing organizational associations between websites can help quickly uncover hidden malicious network structures and enhance attack prevention capabilities. In addition, with the explosive growth of internet information, different websites within the same organization may contain complementary resources. By identifying and integrating data from these related websites, a more comprehensive and efficient knowledge network can be built, promoting information sharing and collaborative analysis [5].

Although organizational-level website management is becoming increasingly important, there is currently limited research addressing this issue. Most existing research focuses on website-level management needs, such as malicious website identification [2, 10]. Those researches mainly analyze various resources of websites, construct an embedded representation of the website using deep learning and other techniques, and use this representation as the input for a classifier to determine whether the target website is malicious.

However, this type of method has two main issues: (1) It focuses on website-level identification tasks. As mentioned above, once a malicious website is discovered, the malicious organizer can easily switch to a new website to continue their activities. Therefore, website-level identification alone cannot fully address such malicious behavior. (2) The various resources required by the method are not always available or accurate. For example, due to CDN technology [12], the IP address associated with the same website may change. Additionally, due to privacy concerns, the WHOIS information of many websites, especially malicious ones, is often incorrect or even nonexistent. The lack of these resources can severely impact the method's accuracy.

To address the above problems, we propose the concept of website association. Website association is an organizational-level alignment task for websites, aiming to determine whether target websites belong to the same organization by analyz-

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_4

ing their resource characteristics. At the same time, we propose a website association model, WebAA, which achieves high-accuracy website association while relying on only a small amount of easily accessible website resources. WebAA consists of four modules that analyze the similarities between target websites using multi-resource and produce the final results through a fusion mechanism. Specifically, in the first module, we first extract the external resources that the target websites rely on. These resources are then categorized into core resources, major resources, and general resources based on their importance, with different weights assigned to each category. Next, inspired by Jaccard similarity, we propose weighted Jaccard similarity and calculate the weighted similarity for each type of resource. Finally, the similarities at each category are combined to produce the overall similarity score S_E based on external resources. In the second module, we first extract and preprocess the HTML text of the target websites. Then, we utilize the BERT model [4] to analyze the similarity between the HTML texts of the target websites, resulting in the similarity score S_T. The third module focuses on the domain name similarity of the websites. We apply the Levenshtein distance algorithm [14] to analyze the domain name similarity between the target websites, resulting in the domain name-based similarity score S_D. In the fourth module, the three similarity scores are weighted to calculate the final similarity score S_F, which is then used to determine whether the target websites belong to the same organization. Extensive experiments on two real-world datasets validate the effectiveness of our model.

In general, the contributions of this paper are as follows:

- **New Concept:** To the best of our knowledge, this is the first time to propose the concept of website association. Website association focuses on the organizations behind websites, enabling regulators to effectively manage legitimate websites and fundamentally combat illicit websites.
- **New Technique:** We innovatively propose a website association model, WebAA, which can accurately and efficiently perform the association task by analyzing only three easily accessible resources of the target websites.
- **New Datasets:** We manually constructed two real-world datasets: one consisting of legitimate websites and their organizational information, and the other comprising illegal websites. We release those datasets to support community researchers in conducting studies related to this field[4].
- **New Promotion:** Extensive experiments on two real-world datasets demonstrate that our model can efficiently associate thousands of website pairs within milliseconds with an accuracy exceeding 90%.

2 Related Work

At present, research on websites mainly focuses on malicious website identification. In this section, we briefly introduce several methods for identifying malicious websites. We also present text matching and similarity calculation tech-

[4] github: https://github.com/SevenZhang123/WebAA-Datasets

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_4

niques based on the BERT model to help readers understand the text similarity calculation method used in our model.

2.1 Malicious Website Identification

Existing methods for identifying malicious websites analyze the characteristics of various resources on the website and use deep learning and other techniques to perform the identification task. IDTracker [14] first divides the target websites into different clusters by Third-party Service IDs. Then constructs a heterogeneous graph to learn the embedding representation of the website based on the association relationship between various resources (domain names, IP addresses, whois records, etc.). And finally uses a classifier to determine whether the target website is a malicious website. Different from IDTracker, DRSDetector [18] uses CBOW, LightGBM and HAN to process different resources of the target website respectively. And finally determines whether the target website is a malicious website through a scoring mechanism. Furthermore, Lei et al. [6] constructed a bipartite graph to capture the interactions between the terminal host and the domain name, the IP resolution structure of the domain name, and the DNS query time series pattern of the domain name. They then used graph embedding techniques to automatically learn the dynamic and discriminative feature representations of the domain name. Based on these feature representations, they predicted whether a newly observed domain name was malicious or benign.

The above methods perform well in identifying malicious websites, but some of the resources they rely on (such as IP addresses, WHOIS information, etc.) are difficult to obtain. Additionally, these methods are focused solely on the website level, which is insufficient for effective website governance. The method we propose in this paper is aimed at the needs of organizational-level website management, which can help reviewers better grasp the website ecology. Additionally, the method relies on only a small amount of easily accessible resource information and can achieve high-accuracy association tasks.

2.2 Text Matching based on BERT Model

The BERT-based text matching task aims to evaluate the semantic similarity or relevance between multiple text fragments. The LTM-B model [7] is a long-text matching method based on BERT. It employs a twin network architecture and adopts a hierarchical approach to divide the text into multiple segments. The BERT model is then used to vectorize the text, generating a matrix representation of the document. Finally, the two document matrices are interacted, pooled, concatenated, and the matching results are output through classification in a fully connected layer. Xia et al. [17] proposed a knowledge-enhanced BERT model, which improved its performance on semantic text matching tasks by directly injecting prior knowledge into BERT's multi-head attention mechanism. The DABERT [15] model takes into account the impact of noise in sentences on model performance, and it is confirmed that subtle noise such as adding, deleting, and changing words in sentences may lead to prediction errors. To solve

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_4

this problem, the authors proposed a novel dual attention method to enhance BERT's ability to capture fine-grained differences in sentence pairs, which to some extent solves the impact of sentence noise on model performance.

The BERT-based text matching method can make full use of the BERT model's ability to extract text information. It adopts a bidirectional attention mechanism and considers the context of the text at the same time, so as to have a deeper understanding of the semantic relationship. This method is suitable for analyzing the semantic structure of long and complex sentences such as HTML text, and has high matching accuracy. In addition, the BERT-based text matching method offers the advantages of flexible transfer learning and task adaptability. It only requires fine-tuning the pre-trained BERT model according to the specific task context to achieve a high matching rate. In this paper, we only use the training dataset to fine-tune the pre-trained BERT model [13] to achieve excellent performance, which greatly reduces the computational cost.

3 WebAA

In this section, we introduce WebAA in detail. The overall architecture of WebAA is shown in Figure 1. WebAA consists of four modules. The first three modules obtain similarity scores based on external resources, HTML texts, and domain names, respectively. Finally, the fourth module integrates them to obtain the final score, which is used to determine whether the target websites belong to the same organization.

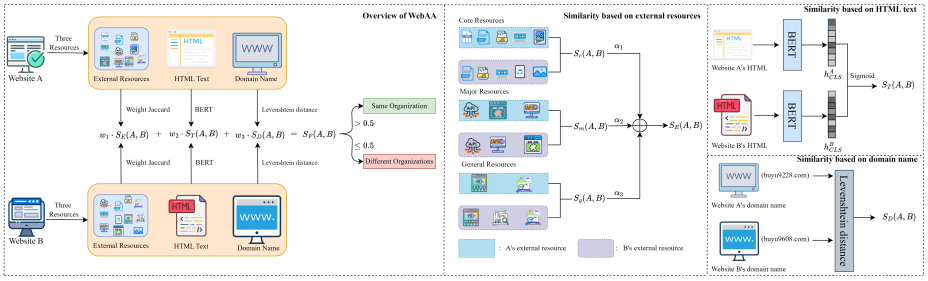

Fig. 1. Overview of WebAA. The left part shows the overall flowchart of the model, while the right part elaborates on the similarity calculation process based on three resources: similarity based on external resources, similarity based on HTML texts, and similarity based on domain names.

3.1 Similarity based on External Resources

External resources are resources loaded by the website during the parsing process, such as CSS files, JavaScript files, images, audio and video, etc. Due to

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_4

storage limitations, websites belonging to the same organization often share a large number of external resources, meaning that the external resources loaded by these websites are hosted on the same server.

Taking target websites A and B as an example, to calculate the similarity score between A and B based on external resources, We first obtain the server address where external resources are stored, denoted as $D_A = \{d_1^A, d_2^A, ..., d_n^A\}$, $D_B = \{d_1^B, d_2^B, ..., d_m^B\}$. This process is completed using BeautifulSoup of the bs4 library. Then, the resources are categorized based on their functional types, including core resources (such as images, audio and video, CSS files, etc.), major resources (such as third-party plug-ins), and general resources (such as advertising and analytics tools). Different weights are assigned to resources of different categorized (core resources have a weight of 3, major resources have a weight of 2, and general resources have a weight of 1), expressed as $W_A = \{w_1^A, w_2^A, ..., w_n^A\}$, $W_B = \{w_1^B, w_2^B, ..., w_m^B\}$. Next, inspired by Jaccard similarity, we propose weighted Jaccard similarity that first calculates the weighted similarity of resources at each category. Taking core resources as an example:

$$S_{\mathrm{c}}\left(A, B\right) = \frac{\sum_{x \in D_A^{\mathrm{core}} \cap D_B^{\mathrm{core}}} \min\left(w_x^A, w_x^B\right)}{\sum_{x \in D_A^{\mathrm{core}} \cup D_B^{\mathrm{core}}} \max\left(w_x^A, w_x^B\right)} \tag{1}$$

Finally, the similarity scores across all categories are integrated to derive the external resources-based similarity score S_E.

$$S_E(A, B) = \alpha_1 S_c(A, B) + \alpha_2 S_m(A, B) + \alpha_3 S_g(A, B) \tag{2}$$

Where, α_1, α_2 and α_3 are trainable weight parameters, and we limit $\alpha_1 + \alpha_2 + \alpha_3 = 1$.

3.2 Similarity based on HTML Texts

We found that although websites under the same organization may serve different functions, their HTML texts show significant similarities. Based on this observation, we calculated the HTML text similarity scores between target websites using the BERT model. BERT has excellent performance in processing long texts and can easily perform transfer learning. For the pre-trained BERT model, only simple fine-tuning on our dataset is required to achieve excellent results.

To calculate the similarity of the HTML texts between target websites A and B, we first use a crawler tool to retrieve the HTML source code of the target websites. Then, we clean the data by removing tags and irrelevant content (such as scripts, style sheets, and other non-essential elements) to obtain clean and meaningful HTML text for input to the BERT model. The input text is first processed by a tokenizer, divided into sub-word units, and mapped into the embedding space. Assuming the input sequence is T, BERT maps it to the initial embedding matrix E : $E = [E_1, E_2, \ldots, E_n]$, where E_i represents the embedding vector of the i-th word. The input embedding is then processed by a multi-layer Transformer encoder. Transformer uses a self-attention mechanism to

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_4

capture contextual information and model the global dependencies of the input sequence:

$$H^l = Transformer(H^{l-1}) \tag{3}$$

Where, H^l represents the embedding matrix of the l-th layer, which contains contextual semantic information. The initial state $H^0 = E$.

After multi-layer Transformer encoding, the embedding representation h_{CLS} of the [CLS] tag is extracted from the output matrix of the final layer as the global semantic feature of the input sequence:

$$H = \text{BERT}(T) \quad \Rightarrow \quad h_{\text{CLS}} = H[0] \tag{4}$$

Where, h_{CLS} represents the HTML text embedding vector of the target website. For target websites A and B, after obtaining the corresponding embedding vectors h^A_{CLS} and h^B_{CLS}, we calculate the similarity score $S_T(A, B)$ through the sigmoid function:

$$S_T(A, B) = \sigma(W \cdot [h^A_{CLS} \oplus H^B_{cls}] + b) \tag{5}$$

Where, $\sigma(x) = \frac{1}{1+e^{-x}}$ is sigmoid function, W and b are trainable parameters, and $\oplus$ represents the embedding vector concatenation operation.

3.3 Similarity based on Domain Names

We observed that the domain names of websites under the same organization are very similar, so we use the domain name similarity score as one of the features to evaluate whether the websites belong to the same organization. In this section, we compute domain name similarity based on the Levenshtein distance. Specifically, we first calculate the Levenshtein distance $lev_{d_A,d_B}(|d_A|, |d_B|)$ between the domain names d_A and d_B of target websites A and B:

$$\text{lev}_{d_A,d_B}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0 \\ \min \begin{cases} \text{lev}_{d_A,d_B}(i-1,j) + 1 \\ \text{lev}_{d_A,d_B}(i,j-1) + 1 \\ \text{lev}_{d_A,d_B}(i-1,j-1) + 1_{(d_{A_i} \neq d_{B_j})} \end{cases} & \text{otherwise.} \end{cases} \tag{6}$$

Where, $1_{(d_{A_i} \neq d_{B_j})}$ is an indicator function, which is equal to 0 when $d_{A_j} = d_{B_j}$ and 1 otherwise.

Since the Levenshtein distance is always an integer greater than or equal to 0, we get the domain name-based similarity score $S_D(A, B)$ by adding 1 to the Levenshtein distance and taking the inverse:

$$S_D(A, B) = \frac{1}{lev_{d_A,d_B}(|d_A|, |d_B|) + 1} \tag{7}$$

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_4

3.4 Fusion Module

In the first three modules, we obtained similarities based on external resources, HTML texts, and domain names. In this module, we fuse these similarities using weighted averaging to obtain the final similarity score $S_F(A, B)$ between the target websites:

$$S_F(A, B) = w_1 \cdot S_E(A, B) + w_2 \cdot S_T(A, B) + w_3 \cdot S_D(A, B) \tag{8}$$

Where, w_1, w_2 and w_3 are trainable parameters, represent the weights of the three similarities. $S_F(A, B)$ represents the final similarity score. If $S_F(A, B) > 0.5$, the target websites are judged to belong to the same organization. If $S_F(A, B) \leq 0.5$, the target websites are judged to belong to different organizations.

4 Experiment

In this section, we conduct extensive experiments aimed at answering the following three research questions:

- **(RQ1)** In the real dataset, are there correlations between websites under the same organization in terms of external resources **(RQ1-1)**, HTML text **(RQ1-2)**, and domain names **(RQ1-3)**? Additionally, are there differences in these resources between websites of different organizations?
- **(RQ2)** How does WebAA model perform on real datasets?
- **(RQ3)** How do different modules of the model (such as similarity of external resources, similarity of html text, and similarity of domain names) affect the model performance?

We first describe the dataset composition, experimental environment, and model hyperparameter settings in Section 4.1. Then, we address the three questions raised in this section through the experiments presented in Sections 4.2 and 4.3.

4.1 Experiment Design

Dataset To verify the effectiveness of our model, we manually constructed two datasets based on real-world data. One is the legal website dataset D_{Legal}, and the other is the illegal website dataset $D_{Illegal}$ (including pornographic, gambling, drug, phishing websites, etc.). The detailed information of two dataset is shown in Table 1.

When building D_{Legal}, we used the website registration number as a reference. The registration number is a unique identifier assigned to an organization in mainland China when it registers a website or application. For multiple websites under the same organization, suffixes such as '-1' and '-2' are added to the registration number to distinguish them. Thus, the registration number serves as the identifier of the website owner. If two websites share the same registration number, they can be confirmed to belong to the same organization. When

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_4

Table 1. Overview of Datasets

Dataset	Web Num	Org Num	Largest Org	Smallest Org
D_{Legal}	4,746	3,127	281	1
$D_{Illegal}$	8,998	1,606	546	1

constructing the dataset, we first downloaded 51,612 Chinese websites from chinaz [3]. Next, we obtained the registration number for each website. Since the registration number information of some websites could not be found, we retained only those with available registration numbers. Finally, we labeled the websites based on their registration numbers and constructed the groundtruth. After screening, we retained a total of 4,746 websites belonging to 3,127 organizations. Among these, the largest organization included 281 websites, while the smallest organization consisted of just one website.

When constructing $D_{Illegal}$, we could not use the registration number as the basis because illegal websites generally do not contain registration number information. Fortunately, Wang et al. [14] published a dataset of illegal websites in their paper. The dataset contains 18,201 websites, which are divided into organizations. Since the life cycle of illegal websites is generally short, we further analyzed the activity of all websites in the dataset and retained only the currently active websites to construct $D_{Illegal}$. The dataset contains a total of 8,998 websites belonging to 1,606 organizations, among which the largest organization contains 546 websites and the smallest organization contains one website.

D_{Legal} and $D_{Illegal}$ represent legal and illegal website scenarios, respectively. By conducting experiments on these two datasets, we can evaluate the model's performance in different scenarios and further demonstrate its wide applicability and robustness across diverse scenarios.

Experimental environment and hyperparameters When training the model, we set the ratio of the training set, validation set, and test set to 4:3:3. The number of epochs to 100, the learning rate to 0.001. We use the cross-entropy loss function to calculate the loss between the model's output and the true labels, and employ the Adam optimizer to compute gradients and update the parameters. In addition, to balance computational cost and performance, we set the embedding vector of the HTML text to 128 dimensions. We use PyTorch to implement WebAA and conduct experiments on two A100 GPUs with 80GB of memory, the server operating system is Ubuntu 20.04.

4.2 Website Resource Analysis (answer RQ1)

In this section, we analyze the correlations and differences in external resources, HTML texts, and domain names between websites belonging to the same organization and those of different organizations to answer RQ1.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_4

External resource analysis (answer RQ1-1) In Section 3.1, we mentioned that websites belonging to the same organization share a significant number of external resources, whereas websites from different organizations do not. To verify this hypothesis, we conducted an experiment on the D_{Legal} dataset. We randomly selected two organizations for analysis, and the experimental results are presented in Figure 2.

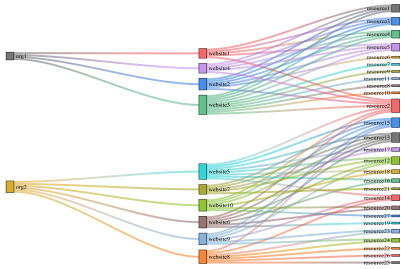

Fig. 2. External resources relied upon by websites under Org1 and Org2.

Considering privacy issues, we have anonymized the organization name, website, and external resource information. The experimental results indicate that websites within the same organization reuse a significant amount of external resources, whereas only a small number of resources are shared between websites from different organizations. For instance, as shown in the figure, only 'resource2' is shared between the websites of org1 and org2. Furthermore, we quantified the external resource reuse rate. Analysis of the two datasets revealed that websites within the same organization had a reuse rate of 66.4%, whereas websites from different organizations had a reuse rate of only 3.8%. These results further validate the accuracy of our idea and the effectiveness of this method for website association analysis.

HTML text analysis (answer RQ1-2) As mentioned in Section 3.2, while websites within the same organization may serve different functions, their HTML texts exhibit significant similarities. To validate this, we conducted experiments on two datasets. Specifically, we selected five organizations from each dataset. Using the BERT model, we generated embedding vectors for the websites' HTML texts and used these vectors to represent them. To intuitively analyze the experimental results, we visualized the embedding vectors, as shown in Figure 3. Since the vector dimension is 128, we used t-SNE [8] to reduce the dimension for easy display.

As illustrated in Figure 3, the HTML text embeddings of websites belonging to the same organization tend to cluster together, while the HTML text embeddings of websites from different organizations are generally farther apart.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_4

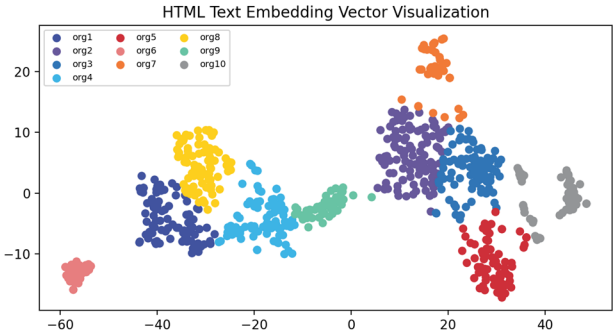

Fig. 3. Visual distribution of HTML text embedding vectors generated based on BERT in 2D space.

This experimental result supports our idea: the HTML texts of websites within the same organization exhibit significant similarities, while those from different organizations are markedly distinct.

Domain name analysis (answer RQ1-3) As mentioned in Section 3.3, the domain names of websites under the same organization are very similar, while those of websites under different organizations are quite different. To verify this, we conducted experiments using two datasets. Specifically, we randomly selected two organizations from each dataset and analyzed the domain name composition of each organization's websites to confirm the validity of our idea.

First, for each organization, we counted the overall distribution of characters in the domain names of all its websites and analyzed the organization-level character distribution. The experimental results are presented in Figure 4. Among them, Org1 and Org2 are organizations in D_{Legal}, Org3 and Org4 are organizations in $D_{Illegal}$.

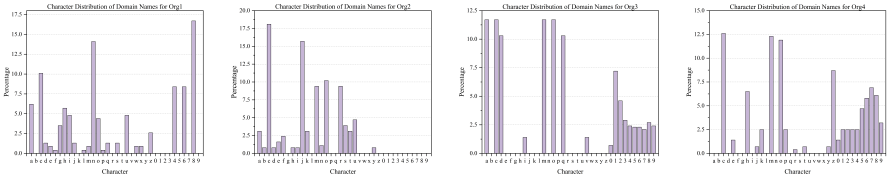

Fig. 4. Overall distribution of characters in the domain names for each organization.

The experimental results show that there are significant differences in character distribution between different organizations. For example, domain names in the Org1 tend to use characters such as 'n' and '8,' while the Org2 does not

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_4

use numbers in domain names at all. Additionally, compared to legitimate websites, illegal websites are more likely to use numbers when constructing domain names. This phenomenon has also been confirmed by Antonakakis et al. [1].

Next, to further investigate the correlation in character distribution of domain names among websites within the same organization, we analyzed those distribution under four organizations. The experimental results are shown in Figure 5. Also for privacy reasons, we hide the real domain name of the website and use *domainX* instead.

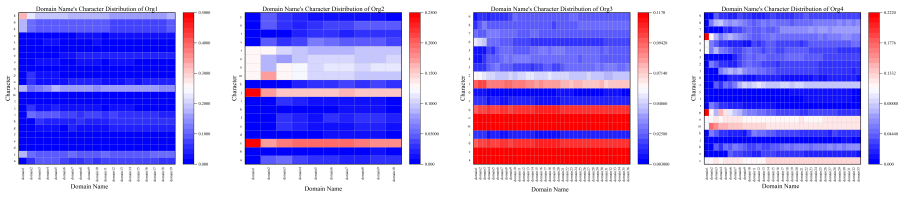

Fig. 5. Correlation heatmap of character distribution in domain names between websites.

The experimental results show that the character distribution of domain names within the same organization is very similar. For example, in Org2, all websites frequently use 'c' and 'j' in their domain names. These results further confirm the validity of our ideas and provide theoretical support for the effectiveness of our model.

4.3 Model Performance

In this section, we answer RQ2 and RQ3 by evaluating the performance of WebAA on two datasets, as well as the performance of each module in WebAA through ablation experiments.

WebAA performance (answer RQ2) In this section, we evaluate the performance of the WebAA model using two real-world datasets. Specifically, we randomly selected 2,000 pairs of websites (1,000 pairs of positive samples and 1,000 pairs of negative samples) from the D_{Legal} and $D_{Illegal}$ datasets, respectively, to form the test samples. We employed four metrics—accuracy, recall, F1 score, and the ROC curve—to quantitatively assess the model's performance.

The experimental results are presented in Table 2 and Figure 6. On both datasets, the accuracy, recall and F1 score of the WebAA model exceed 90%. Specifically, on the D_{Legal} dataset, these three metrics surpass 95%, highlighting the exceptional performance of the WebAA model in the website association task. In addition, the ROC curve in Figure 6 shows that at a lower false positive rate, the model is able to achieve a higher true positive rate, thus being able to effectively perform the website association task. However, compared to

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_4

Table 2. Experimental results on two datasets

Model	D_{Legal}				$D_{Illegal}$			
	Acc	Recall	F1	Time(s)	Acc	Recall	F1	Time(s)
WebAA	**97.15**	95.10	**97.09**	0.45	**92.95**	**92.70**	**92.93**	0.45
WebAA$_r$	91.75	83.70	91.03	0.03	86.90	75.80	85.26	**0.12**
WebAA$_h$	96.80	96.50	96.79	0.38	90.85	88.50	90.63	0.40
WebAA$_d$	82.45	**97.30**	84.72	**0.02**	69.80	89.80	74.83	0.14

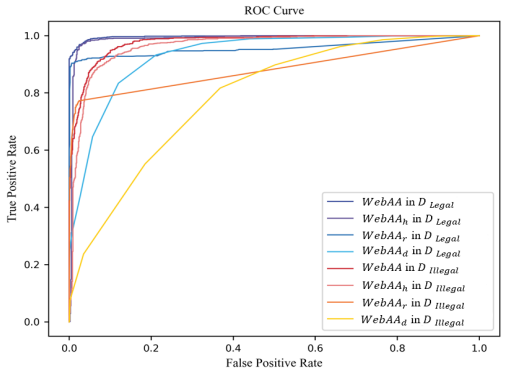

Fig. 6. ROC curves of four models on two datasets.

the D_{Legal} dataset, the model performed slightly worse on the $D_{Illegal}$ dataset. We attribute this to deliberate dissimilarity operations performed by the organizations behind illegal websites to avoid detection. These operations include modifying HTML text, reducing domain name similarity, and minimizing the overlap of external resources. Such modifications introduced slight deviations in our model's performance during website association tasks. Nevertheless, it is worth noting that despite these dissimilarity operations, our model still achieved over 90% on all three metrics, which indirectly demonstrates its robustness and resistance to interference. In terms of time cost, the WebAA model can complete the association of thousands of website pairs within milliseconds on both datasets, fully demonstrating its efficiency and practicality.

Ablation experiment (answer RQ3) In this section, we evaluate the impact of different modules through ablation experiments. Specifically, we design three models: WebAA$_r$ (considers only external resource similarity), WebAA$_h$ (focuses solely on HTML text similarity), and WebAA$_d$ (examines only domain name similarity) to analyze the contribution of each module to the website association task. We apply the same samples and evaluation metrics used for WebAA to assess the performance of these three models. The experimental results are presented in Table 2 and Figure 6.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_4

The experimental results indicate that the performance of different modules varies significantly. Among them, the WebAA$_h$ model demonstrates the best performance, with each metric being only about 1% lower than that of the WebAA. This suggests that, in the website association task, the similarity of HTML text between target websites is a crucial factor. The performance of the WebAA$_d$ model is relatively poor, with an accuracy of only 69.80% on $D_{Illegal}$. We analyzed that this is because, compared to HTML text and external resources, domain names are simply identifiers for websites and are more prone to dissimilarity. Therefore, using domain similarity alone will not yield the desired results. However, compared to other models, WebAA$_d$ has a higher recall rate and can efficiently recall positive samples.

In general, in the website association task, the three modules each have their own advantages and contribute to the WebAA model to varying degrees. Ultimately, the fusion of these modules enhances the overall performance of the WebAA model.

5 Conclusion

In this paper, we introduce the concept of website association and propose the first model, WebAA, for this task. This model determines whether target websites belong to the same organization by analyzing the similarity of multiple resources. As a further extension of website identification, website association focuses on the organization behind the website, providing a new technical approach for mapping cyberspace ecology and combating illegal websites. Extensive experiments on two manually constructed datasets demonstrate that our model can efficiently associate thousands of website pairs within milliseconds with excellent performance. This fully verifies its significant advantages in terms of accuracy, efficiency, and robustness. Additionally, we will open-source the relevant datasets to assist researchers in the community with their work in this field.

Acknowledgments. This work is supported by the Scaling Program of Institute of Information Engineering, CAS (Grant No. E3Z0041101).

References

1. Antonakakis, M., Perdisci, R., Nadji, Y., Vasiloglou, N., Abu-Nimeh, S., Lee, W., Dagon, D.: From throw-away traffic to bots: Detecting the rise of dga-based malware. In: Proceedings of the 21th USENIX Security Symposium, 2012. pp. 491–506. USENIX Association (2012)
2. Chen, J., Zheng, S., Cheng, Y., Zhang, Z.: Data mining based analysis of online gambling sites and illicit financial flows. In: 2024 International Conference on Cloud Computing and Big Data, ICCBD 2024. pp. 205–211. ACM (2024)
3. Chinaz: https://top.chinaz.com/all/ (2024)

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_4

4. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019. pp. 4171–4186. Association for Computational Linguistics (2019)
5. Ji, S., Pan, S., Cambria, E., Marttinen, P., Yu, P.S.: A survey on knowledge graphs: Representation, acquisition, and applications. IEEE Trans. Neural Networks Learn. Syst. **33**(2), 494–514 (2022)
6. Lei, K., Fu, Q., Ni, J., Wang, F., Yang, M., Xu, K.: Detecting malicious domains with behavioral modeling and graph embedding. In: 39th IEEE International Conference on Distributed Computing Systems, ICDCS 2019. pp. 601–611. IEEE (2019)
7. Long Liu, Xin Liu, L.C.C.T.: Long text matching model based on bert. Computer system applications (2018)
8. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
9. Sánchez-Rola, I., Dell'Amico, M., Balzarotti, D., Vervier, P., Bilge, L.: Journey to the center of the cookie ecosystem: Unraveling actors' roles and relationships. In: 43rd IEEE Symposium on Security and Privacy, SP 2022. pp. 1990–2004. IEEE (2022)
10. Starov, O., Zhou, Y., Zhang, X., Miramirkhani, N., Nikiforakis, N.: Betrayed by your dashboard: Discovering malicious campaigns via web analytics. In: Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018. pp. 227–236. ACM (2018)
11. Subramani, K., Melicher, W., Starov, O., Vadrevu, P., Perdisci, R.: Phishinpatterns: measuring elicited user interactions at scale on phishing websites. In: Proceedings of the 22nd ACM Internet Measurement Conference, IMC 2022. pp. 589–604. ACM (2022)
12. Subramani, K., Perdisci, R., Skafidas, P., Antonakakis, M.: Discovering and measuring cdns prone to domain fronting. In: Proceedings of the ACM on Web Conference 2024, WWW 2024. pp. 1859–1867. ACM (2024)
13. THUCNews: Bert-chinese-text-classification-pytorch. https://github.com/649453932/Bert-Chinese-Text-Classification-Pytorch
14. Wang, C., Li, Z., Yin, J., Liu, Z., Zhang, Z., Liu, Q.: Idtracker: Discovering illicit website communities via third-party service ids. In: 53rd Annual IEEE/IFIP International Conference on Dependable Systems and Network, DSN 2023. pp. 459–469. IEEE (2023)
15. Wang, S., Liang, D., Song, J., Li, Y., Wu, W.: DABERT: dual attention enhanced BERT for semantic matching. In: Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022. pp. 1645–1654. International Committee on Computational Linguistics (2022)
16. Worldwidewebsize: The size of the world wide web (the internet). https://www.worldwidewebsize.com/ (2025)
17. Xia, T., Wang, Y., Tian, Y., Chang, Y.: Using prior knowledge to guide bert's attention in semantic textual matching tasks. In: WWW '21: The Web Conference 2021. pp. 2466–2475. ACM / IW3C2 (2021)
18. Zhang, Y., Fu, X., Yang, R., Li, Y.: Drsdetector: Detecting gambling websites by multi-level feature fusion. In: IEEE Symposium on Computers and Communications, ISCC 2023. pp. 1441–1447. IEEE (2023)

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_4