An Abnormal External Link Detection Algorithm Based on Multi-Modal Fusion

Zhiqiang Wu, Henan Police College, China*

iD https://orcid.org/0000-0002-5373-3983

ABSTRACT

Website link detection is an important means to ensure the security of the external chain. In the past, it was mainly realized through blacklisting and feature engineering-based machine learning, which has the problems of slow detection speed and weak model generalization ability. The development of neural networks has brought a new solution to the security detection of the external chain of the website. To address the performance bottleneck caused by the variable content length of web pages, this article introduces an innovative approach: a website external link security detection algorithm based on multi-modal fusion. It extracts text, dynamic script, and image features separately, and constructs a deep fusion model that combines these multi-modal features. Compared with the previous research results, the proposed method is superior to the traditional single-mode method, and can quickly and accurately identify malicious web pages. The accuracy and F1 value are improved by 2.7% and 0.026.

KEYWORDS

feature extraction, malicious content, multimodal fusion, security detection, website external links

INTRODUCTION

With the rapid development of information technology and the popularization of the Internet, the number of websites on the Internet has increased exponentially. In order to provide users with richer information resources and promote cooperation and interaction with other websites or institutions, a lot of external links are generally introduced into the website. Due to information updates, domain name changes, hacker attacks, and other reasons, if you link to an insecure external website, it will pose a security risk to users. Such risks can include malicious links, erotic gambling sites, or web pages containing malicious code that may lead to the disclosure of the user's personal information, computer infection, economic losses, and other problems (Tenis & Santhosh, 2021). In addition, if you link to external websites containing harmful information, it will seriously damage the reputation of the organization, and users may doubt the professionalism, trust, and network security capabilities of the organization, which will affect user's access to and use of the organization's website. Therefore, ensuring the security of the external link of the website is crucial for the website.

*Corresponding Author

It is an important means to carry out regular inspections of the external chain of the website to ensure the security of the external chain. However, due to the large number of websites and pages, it is undoubtedly unrealistic for website security managers to use manual inspection. With the development of computer technology, the research on the security detection of external links of websites by computer programs has been widely concerned, and many detection schemes have been proposed by scholars at home and abroad. The earliest detection method used the blacklist technique, which preconstructed a blacklist listing all known harmful domain names. When a user visits a website, they check whether its domain address is in the blacklist to detect harmful external links. This method has the advantage of high detection accuracy, but it needs to ensure the timely maintenance of the black and white list, which has certain limitations and lag and cannot effectively judge the security of unknown web pages (Darwish et al., 2023). To solve this problem, some scholars have proposed a method based on dynamic behavior analysis, which analyzes the behavior of the website host, such as access records, execution processes, etc., to analyze whether the website host has abnormal behavior and find out the abnormal external chain. This method has the ability to detect unknown viruses and malicious codes, but the detection speed is slow because it needs to simulate the running state of malicious web pages and analyze them.

With the development of data mining and machine learning technology, a website off-link security detection method based on machine learning has been proposed (Jerjes et al., 2023; Venugopal et al., 2021). This method has a certain generalization ability, but due to the great impact of the selection of webpage features on the model recognition effect, the workload in the feature engineering stage is relatively large. At the same time, the traditional machine learning technology cannot learn the contextual semantic features of web text, resulting in a certain bottleneck in the recognition effect.

In the past few years, the field of external chain detection has witnessed a shift toward deep learning-based approaches driven by the rapid advancements in machine learning and artificial intelligence technology. According to the existing literature, text features are mostly used, and due to the variable length of Chinese text on web pages (Naim et al., 2023), in order to achieve the feasibility of model training, in addition to short text features such as Uniform Resource Locator(URL) and tags, part of text content from web pages is generally extracted for model training, resulting in poor practicability of the trained model. In addition, with the development of communication technology, a large number of web pages contain not only text information but also a lot of multimedia information, such as pictures, videos, and sounds. It is not good to judge whether a web page has malicious information only through text information. In view of these problems, in this research, the website link security detection is regarded as a binary classification problem. By integrating the features of webpage text, dynamic script, and image, an innovative intelligent detection algorithm for website link security based on multimodal fusion is proposed. The main work of this paper includes:

1. The FastText model is used to extract the text features of web pages, aiming at the problem that the content length of web pages is not fixed.
2. Aiming at the performance bottleneck of web page detection using text alone, this paper introduces an innovative approach: a website external link security detection algorithm based on multimodal fusion. It extracts text, dynamic script, and image features separately and constructs a deep fusion model that combines these multimodal features.
3. Through comparative experiments, the effectiveness of the web security intelligent detection algorithm proposed in this paper based on multimodal feature fusion is verified.

The following organizational structure of this paper is as follows: The Related Work section briefly introduces the related research work; the Methodology section introduces the feature extraction method and the multimodal feature fusion model. The Experiment section discusses the effectiveness of the proposed algorithm verified by comparison experiments. The Discussion section summarizes and looks forward to the work of this paper.

13 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/article/an-abnormal-external-link-detection-algorithm-based-on-multi-modal-fusion/337894

Related Content

Firm Value and Self-Insurance: Evidence from Manufacturers in California
Mu-Sheng Chang, Hsin-Hui Chiuand Yanbo Jin (2019). *International Journal of Risk and Contingency Management (pp. 59-73).*
www.irma-international.org/article/firm-value-and-self-insurance/216869

Hacking: Evolution, Conceptualization, and the Perpetrators
Carolina Roque, Maria Canudo, Samuel Moreiraand Inês Sousa Guedes (2023). *Contemporary Challenges for Cyber Security and Data Privacy (pp. 83-107).*
www.irma-international.org/chapter/hacking/332717

A Comprehensive Review of Privacy Preserving Data Publishing (PPDP) Algorithms for Multiple Sensitive Attributes (MSA)
Veena Gadadand Sowmyarani C. N. (2023). *Information Security and Privacy in Smart Devices: Tools, Methods, and Applications (pp. 142-193).*
www.irma-international.org/chapter/a-comprehensive-review-of-privacy-preserving-data-publishing-ppdp-algorithms-for-multiple-sensitive-attributes-msa/321342

Assurance and Compliance Monitoring Support
Peter Goldschmidt (2001). *Information Security Management: Global Challenges in the New Millennium (pp. 135-154).*
www.irma-international.org/chapter/assurance-compliance-monitoring-support/23365

Information Security Practices in Small-to-Medium Sized Businesses: A Hotspot Analysis
Kent Marettand Tim Barnett (2021). *Research Anthology on Privatizing and Securing Data (pp. 2023-2043).*
www.irma-international.org/chapter/information-security-practices-in-small-to-medium-sized-businesses/280270