International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Semantic Analysis for Unwanted Keyword Filtering in Search History

P K Gokul[1]; Dr. Vaidehi.V[2]

[1]PG Student; [2]Professor
Department of Computer Applications, Dr. M.G.R. Educational and Research Institute, Chennai – 6000 95
[1]polasikumargokul@gmail.com; [2]vaidehi.mca@drmgrdu.ac.in

ABSTRACT:

Efficiently filtering unwanted keywords from search history data demands sophisticated semantic analysis techniques. This paper proposes a method harnessing natural SVM (Support Vector Machine) algorithms to comprehend the contextual nuances and meanings of search queries. Leveraging word embedding models, semantic similarity computation, and topic modelling our approach categorizes search queries based on their semantic relevance and potential undesirability. By dissecting the semantic structure of search queries, our method effectively distinguishes between relevant and unwanted keywords, enhancing search history filtering accuracy. Evaluation using a dataset of search queries showcases the superiority of our semantic analysis-based approach over traditional keyword-based filtering methods. This research contributes to the enhancement of search history data management in applications such as privacy protection and content moderation

Keywords: Web Crawler, SVM (Support Vector Machine),

I.INTRODUCTION:

People's lives now revolve around the Internet, which is now essential. But even as the Internet promotes affluence, it also creates issues such as gambling, pornography, fraudulent medical websites, and illegal websites. The quantity of dangerous websites is rising even after a variety of detection techniques have been used. Internet users' health is at risk due to the abundance of hazardous content, particularly for children and teenagers [1], [2].Scholars have devised numerous techniques, such as machine learning-based approaches and heuristic methods, to identify fraudulent websites. Typically, content masking spam offers authentic content while concealing harmful information that is undetectable to users and browsers, increasing the false-positive rate of detection techniques.

The World Wide Web's worldwide address for documents and other resources is called a URL. As demonstrated in Figure 1 below, a URL is made up of two main parts: the Name of the Resource (which contains the IP address or domain name of the resource) and the Protocol Identifier (which indicates the protocol to be used). These parts are separated by a colon and two forward slashes.

A redirection spam sends consumers to an irrelevant page while providing the crawler with a page with incorrect content. The ideal method for redirecting Web spam is JavaScript redirection, which sends spam content to a crawler that doesn't require scripts and then automatically reroutes a browser that can interpret the script to a different URL when the page loads [3]. One common example of a malicious website employing the redirection spam tactic is http://www.bowas.cn/. From a script-agnostic standpoint, it offers gossip and entertainment. To trick the parser, the encoded universal resource location (URL) is given as a string concatenation rather than a single string.

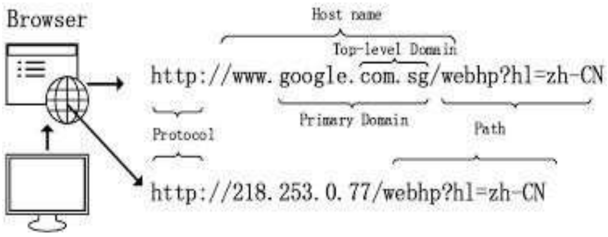

Browser
Host name
Top-level Domain
http://www.google.com.sg/webhp?hl=zh-CN
Protocol
Primary Domain
Path
http://218.253.0.77/webhp?hl=zh-CN

Figure 1: Components of a URL

The browser only displays a portion of the material in content-hiding spam, which includes text and hyperlink-hiding spam, as opposed to redirection spam and hidden I Frame spam, which provides the crawler and browser access to the same page. Put differently, readers cannot see certain content, including the anchor text. One common type of content-concealing spam is hyperlink-hiding spam, which is thoroughly classified by Geng et al. [4]. According to the data, a large number of trustworthy websites have hyperlinks and hidden anchors. Regretfully, though, the majority of concealed anchors also include offensive language. Consequently, a trustworthy website with concealed linkages and textThe prevalence of website design tools, particularly JavaScript, facilitates the use of I Frame, content concealment, and malicious redirection. It also makes it more challenging to identify rogue websites. Currently, one of the biggest obstacles to malicious website detection is the difficulty in figuring out whether or not the parsed online content is suitable for study. It causes malicious website detection techniques to work poorly.

Even though it appears impossible to overcome, there is still a method to bypass the complicated resolution procedure and all the confusion created by spam tactics. The best method to evaluate a website is to look at what it shows visitors in the end. Put differently, if consumers eventually discover fraudulent medication website. Thus, based on the aforementioned study, we take the perspective of the end-user and suggest a technique for detecting harmful websites that uses a Convolutional Neural Network (CNN) to identify and learn from screenshot photos. Our technique realizes the independence of the previously required information supplied by the source code of parsed webpages, hence removing the effect of chaotic and complicated information. Additionally, it makes use of CNNs' strong performance in picture learning and classification. The suggested approach can successfully stop the three spam strategies listed above by lowering the false positive rate and raising the coverage rate.

II. LITERATURE SURVEY

According to E. Sorio, A. Bartoli, et al., (2013) Internet security issues sometimes entail the fraudulent alteration of a website, such as the installation of new pages at URLs where none should exist. Detecting the presence of such hidden URLs is extremely challenging because they do not surface during normal browsing and are typically not indexed by search engines. Most crucially, drive-by attacks that direct users to hidden URLs, such as phishing credentials, may mislead even technically competent users since such hidden URLs are increasingly located within trustworthy sites, rendering HTTPS authentication useless. In this paper, we offer a method for recognizing such URLs based only on their lexical properties, allowing us to inform the user.

According to N. Singh, J. Kumar, A.K. Singh and A Mohan, et al., (2022) In cloud computing, data outsourcing to the cloud is becoming more and more commonplace. A dynamic design that closes gaps and prevents intrusions into the cloud environment is necessary. The newest and most successful method for successfully retrieving the encrypted documents from the cloud is multi-keyword fuzzy search. This method is used in the proposed research project in conjunction with the n-gram corpus algorithm to help achieve sub-linear search time and a relevance score. This algorithmic combination can forecast the recovered data's privacy rating metric. Simulation study has been carried out to maintain user-oriented privacy in a cloud environment in order to validate the opposing strategy. The suggested method's ranking precision is contrasted with that of Fu et al.

According to G. Nath and G. Adhi, et al., (2019) WhatsApp now has 1.5 billion users worldwide. Some individuals have exploited the site to disseminate rumours and fraudulent messages, despite its intended purpose of connecting with family, friends, and coworkers. When a WhatsApp user receives a forward message, they have limited tools to check its authenticity. To check, the user can contact the forwarder or utilize Google search. Verifying the authenticity of WhatsApp messages with a Google search is time-consuming, and users cannot rely on Google for every message they receive. We will analyse bogus material using natural language processing and supervised machine learning to detect trends and discriminate between real and fraudulent communications. We'll also expand our exploration.

III. PROPOSED METHOD

One such frank system limited in its capacity for relearning is the improved Convolution Neural Network (CNN) model put out by the authors. The inventors of this dynamic system have included a character-level embedding layer ahead of the convolution layer, enabling the system to understand the inherent relationships between the characters in the query string. The CNN filters that might extract the query string's fine-grained properties have also been updated by the creators of this system. When compared to Random Forest (RF) and Support Vector Machine (SVM), the test results demonstrate that the current model has a reduced False Positive Rare (FPR).

Current methods include message content analysis (e.g., embedded URL analysis and message clustering) and account profile analysis. Because account profiles contain the original information of ordinary users, which spammers are likely to keep intact, account profile analysis is seldom useful for identifying compromised accounts. By taking advantage of the trusted relationships and well- established connections between the genuine account owners and their friends, malicious actors may effectively disseminate malware, phishing links, or spam advertisements without getting stopped by the service providers.

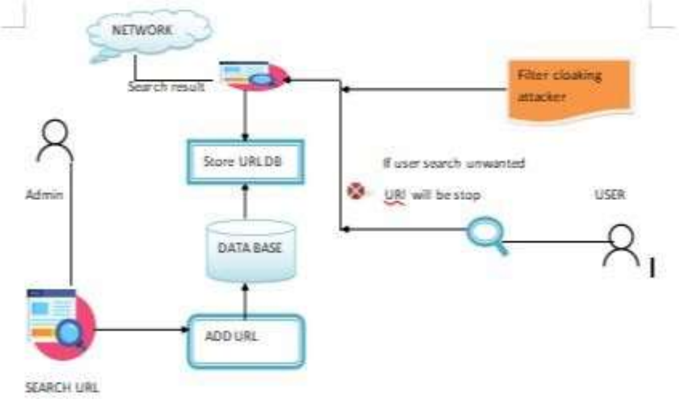


Fig: The System Architecture

IV. EXPLANATION

Admin:

The admin component is responsible for adding new URLs to the system and for viewing the search results.

Database:

The database stores the URLs that have been added to the system and the results of the search for each URL.

Users:

The user component allows users to search for URLs and to view the results of the search

The system works follows:

a. The admin adds a new URL to the system.

b. The system searches for the URL on the internet.

c. The system stores the results of the search in the database.

d. A user searches for a URL.

e. The system compares the user's search query to the URLs in the database.

f. If the system finds a match, it returns the results of the search to the user.

g. If the system does not find a match, it returns a message to the user indicating that the URL is not in the system.

The system also includes a filter for cloaking attacks. A cloaking attack is a type of attack in which a malicious website is disguised as a legitimate website. The system filters out cloaking attacks by checking the IP address of the website that the user is searching for. If the IP address is not in the list of trusted IP addresses, the system returns a message to the user indicating that the website is not trusted.

The system is intended to be precise and efficient. The system makes use of a distributed database and caches search results among other strategies to increase efficiency. The system employs a range of strategies to increase its accuracy, including the use of several search engines and different algorithms to weed out harmful websites.

V. MODULES

Admin:

The "Login" option allows users to access the system by entering their credentials such as username and password. Once logged in, users can utilize the "Add New URL" feature to input a new URL along with an optional description. The "View All Results" functionality displays a list of all URLs added by the user, providing options to edit or delete each entry as needed. Users can also use the "View Blocked URL" option to see a list of URLs that have been blocked, with the ability to unblock them if necessary. Finally, the "Logout" option securely ends the user's session, ensuring privacy and security. These features collectively facilitate user interaction and management within the system.

User:

The given options pertain to a software or platform where users can engage in specific actions. "Register" refers to creating an account or joining the platform, providing necessary information. "Login" allows existing users to access their accounts securely. "Search" enables users to look for specific information, content, or users within the platform. "View result" likely pertains to displaying search outcomes or other relevant data based on user input. Finally, "Exit" allows users to log out or leave the platform/application. These options collectively facilitate user interaction, navigation, and functionality within the system.

Attacker:

The provided sequence of actions "Login - Add Malicious - Logout" suggests a malicious intent within a system or platform. First, the individual logs into a system, gaining initial access likely through legitimate means such as using valid credentials. Next, they perform an unauthorized action by adding something malicious, which could refer to inserting harmful code, uploading infected files, or manipulating data to compromise security or cause harm. Finally, they log out, potentially to cover their tracks and avoid detection. This sequence highlights the importance of robust security measures such as multi-factor authentication, access controls, and regular security audits to detect and prevent such malicious activities

VI. RESULT & DISCUSSION

Admin Login: This action allows a user to access their account by providing their credentials, typically a username and password. Upon successful login, the user gains access to specific features or content based on their role or permissions**.**

Add New URL: Users with appropriate privileges can add new URLs to a database or system. This feature is common in systems where users can bookmark or save URLs for later reference Oshariss with others.

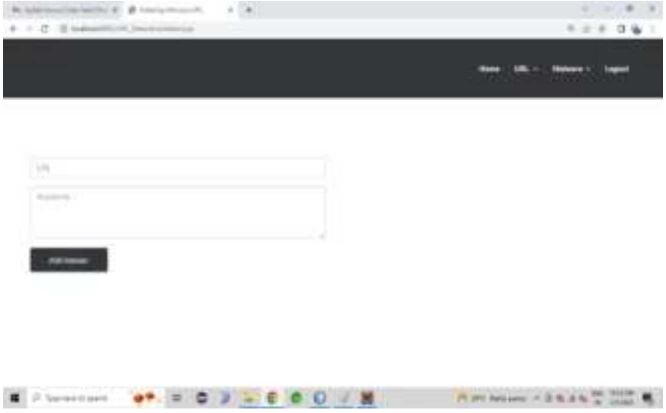
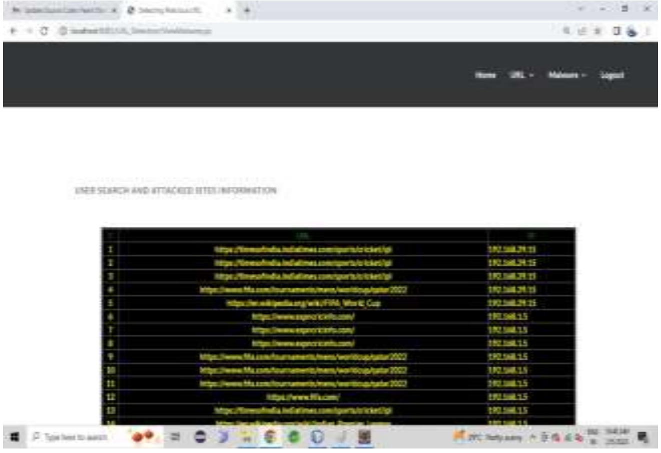


Fig 1: Admin add new URL page.

View All Result: This action displays all the URLs stored in the system, along with any associated information such as descriptions or tags. Users can use this feature to browse through their saved URLs or search for specific ones.



USER SEARCH AND ATTACKED SITES INFORMATION

[illegible]	URL	IP
1	https://timesofindia.indiatimes.com/sports/cricket/ipl	192.168.29.15
2	https://timesofindia.indiatimes.com/sports/cricket/ipl	192.168.29.15
3	https://timesofindia.indiatimes.com/sports/cricket/ipl	192.168.29.15
4	https://www.fifa.com/tournaments/mens/worldcup/qatar2022	192.168.29.15
5	https://en.wikipedia.org/wiki/FIFA_World_Cup	192.168.29.15
6	https://www.espncricinfo.com/	192.168.1.5
7	https://www.espncricinfo.com/	192.168.1.5
8	https://www.espncricinfo.com/	192.168.1.5
9	https://www.fifa.com/tournaments/mens/worldcup/qatar2022	192.168.1.5
10	https://www.fifa.com/tournaments/mens/worldcup/qatar2022	192.168.1.5
11	https://www.fifa.com/tournaments/mens/worldcup/qatar2022	192.168.1.5
12	https://www.fifa.com/	192.168.1.5
13	https://timesofindia.indiatimes.com/sports/cricket/ipl	192.168.1.5
14	[illegible]	192.168.1.5

Fig 2: Admin View the URL page.

View Blocked URL: Some systems may have a feature to block certain URLs for various reasons such as security concerns or content filtering. Users with appropriate permissions can view a list of blocked URLs and possibly take action to unblock them if necessary.

Suspicious Web Page Blocked

Fig 3: Admin block the URL page.

Attackers:

Login: Logging in typically refers to the process of gaining access to a computer system, network, or application by providing valid credentials such as a username and password. It establishes a secure session that allows the user to interact with the system or application.

Add Malicious URl:

The term "Add Malicious" is concerning as it suggests an intent to introduce harmful or malicious content into a system or application. Malicious content can come in various forms such as malware, viruses, or exploits designed to compromise security, steal data, or disrupt normal operations. Intentionally adding malicious content is unethical and often illegal.

Fig: Attackers Add URL page.

Attackers View the URLs: Viewing the URL in the context of an attacker involves examining the structure and parameters for potential vulnerabilities, such as injection attacks or parameter tampering, to exploit weaknesses and gain unauthorized access or manipulate system behaviour. Attackers analyse URL components like query strings, cookies, and hidden fields to identify and exploit security flaws in web applications or services.

Logout:

Ending the session after performing the unauthorized action

User:

Register:

Registering typically refers to creating a new account or profile within a system or application. This process involves providing necessary information such as username, password, email address, and possibly additional details depending on the platform. Registration is crucial for users to access personalized features and functionalities.

Login:

Logging in involves using previously registered credentials (like username and password) to gain access to a system or application. Successful login verifies the user's identity and grants appropriate permissions based on their account privileges

Search:

Searching usually involves querying a database or repository for specific information based on user-provided criteria. This action can occur in various contexts such as searching for products on an e-commerce website, finding articles in a knowledge base, or looking up user profiles in a social networking platform.

Fig 1: User Search page.

View Result:

Viewing results refers to the presentation of information retrieved from the search or query performed earlier. Users typically interact with these results to obtain relevant data, explore content, or make informed decisions based on the displayed information.

Fig 2: User View The URLs.

Exit:

Exiting signifies ending the current session or application instance. It can involve logging out of a user account, closing a program or app window, or terminating a session on a website or platform.

Fig 3: View the Result Page.

VII. CONCLUSION

The content, functionality, and style of mobile web pages vary dramatically from those of desktop websites. As a result, conventional approaches for detecting fraudulent activity that rely on static aspects of desktop web pages do not perform effectively on mobile-specific pages. We conceived and developed Kayo, a quick and reliable static analysis approach for detecting mobile harmful websites. Kayo makes these detections by analyzing 44 mobile-relevant characteristics from webpages, 11 of which are newly uncovered mobile-specific features. Kayo achieves 90% classification accuracy and finds several dangerous mobile webpages in the wild that are not discovered by existing approaches like Google Safe Browsing and Virus Total. Finally, we created a browser plugin with Kayo that gives users real-time feedback. We infer that Kayo detects. We conclude that kayo discovers new mobile-specific dangers, such as websites hosting known fraud numbers, and so represents the first step in identifying new security concerns in the current mobile web., FUTURE WORK: KAYO recognized harmful webpages in the wild that were not detected by previous approaches. We thoroughly study these web pages before discussing kayo's limits and future development.

VIII. REFERENCES

[1] F. D. Abdi, and L. Wenjuan, "Malicious url detection using convolutional neural network," Journal International Journal of Computer Science, Engineering and Information Technology, vol. 7, No. 6, 2017, pp. 1–8.

[2] S.Thakur, E. Meenakshi, and A. Priya, "Detection of malicious URLs in big data using RIPPER algorithm," 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology, (RTEICT), IEEE, 2017, pp. 1296-1301.

[3] V.D. Naveen, K. Manamohana, and R. Verma, "Detection of Malicious URLs using Machine Learning Techniques", International Journal of Innovative Technology and Exploring Engineering,Vol.8, 2019, pp. 389-393

[4] D. Sahoo, C.Liu, and C.H. Steven, "Malicious URL Detection using Machine Learning: A Survey", eprint arXiv: 1701.07179. (2017)

[5] M.S.I. Mamun, M.A. Rathore, A.H. Lashkari, N. Stakhanova, and A.A. Ghorbani, "Detecting malicious URL using lexical analysis," In International Conference on Network and System Security, Springer, Cha, 2016, pp. 467-482.

[6] F. Vanhoenshoven, G.Nápoles, R. Falcon, K. Vanhoof, and M. Köppen, "Detecting malicious URLs using machine learning techniques. In 2016 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, 2016, pp.1-8.

[7] W. Kong and J. Allan, "Extending faceted search to the general web", In Proceeding on ACM International Conference on Information Knowledge Management, 2014, pp. 839–848.

[8] K. Balog, E. Meij and M. De Rijke, "Entity search: Building bridges between two worlds", In Proceeding on 3rd International Semantic Search Workshop, 2010, pp. 1–9.

[9] C. Li, N. Yan, S.B. Roy, L. Lisham and G. Das, "Facetedpedia: Dynamic generation of query-dependent faceted interfaces for Wikipedia", in Proceeding on 19th International Conference World Wide Web, 2010, pp. 651–660.

[10] W. Dakka and P.G. Ipeirotis, "Automatic extraction of useful facet hierarchies from text databases", In Proceeding on IEEE 24th International Conference Data Engineering., 2008, pp. 466–475.s