Web Filtering Software for Controlled Internet Access

[1]PrathmeshNavale,[2]VinayakDesai,[3]OmTakale,[4]VinodGaddam,
[5]DiptiP.Pandit,[6]Gauriv.Ghule
[123456]DepartmentofE&TCEngineering,
[1234]VishwakarmaInstituteofInformationTechnology(VIIT), Pune, India
[56]VishwakarmaInstituteofTechnology(VIT), Pune, India
Emails: prathmesh.22420235@viit.ac.in, vinayak.22420144@viit.ac.in, om.22420092@viit.ac.in, vinod.22420255@viit.ac.in
dipti.pandit@vit.edu, gauri.ghule@vit.edu

***Abstract*—The rapid increase in internet usage across educational institutions and corporate organizations has created significant challenges related to productivity, bandwidth management, and cybersecurity. While internet access is essential for research, communication, and operations, unrestricted browsing often results in misuse of resources and exposure to security threats. Traditional web filtering systems mainly depend on static URL blacklists and DNS-based blocking, which are no longer sufficient to handle dynamic web content, encrypted HTTPS traffic, and constantly changing domain names. This paper presents an intelligent web filtering framework that combines keyword-based content analysis, structural HTML inspection, domain reputation scoring, and statistical performance evaluation to improve filtering accuracy. The system is implemented using a proxy-based architecture that intercepts user requests, analyzes web content in real time, classifies websites into predefined categories, and applies policy-based access control. In addition, process capability indices (Cp and Cpk) are used to evaluate system stability and response time consistency, ensuring reliable performance. Experimental results show improved classification accuracy, lower false-positive rates, and better bandwidth optimization compared to traditional filtering approaches. The proposed framework offers a scalable, adaptive, and practical solution for deployment in colleges and corporate network environments.**

***Index Terms*—Web Filtering, Content Analysis, Structural Analysis, Proxy Server, Network Security, Process Capability, Access Control.**

I. Introduction

Internet access has become an essential resource for Internet connectivity has become a fundamental requirement in both academic institutions and corporate organizations. Colleges rely on online platforms for research, learning management systems, communication, and collaboration. Similarly, companies depend on internet-based tools for business operations, data exchange, and cloud services. However, unrestricted access to the internet can negatively impact productivity and network performance. Websites such as social media platforms, video streaming services, gaming portals, and other non-work-related sites often consume excessive bandwidth and reduce employee or student focus. In addition, access to malicious or compromised websites increases the risk of malware attacks, phishing, and data breaches. Most conventional web filtering solutions rely on static blacklists or DNS-level filtering techniques. While these approaches are simple to implement, they have several limitations. They cannot effectively block newly created domains, dynamically generated URLs, or encrypted HTTPS traffic. Furthermore, simple keyword-based blocking often results in high false-positive rates because it does not consider contextual meaning or structural patterns within web pages. To address these limitations, there is a growing need for a more intelligent and adaptive web filtering framework. Such a system should analyze not only textual content but also the structural components of web pages to ensure accurate classification and reliable access control. This paper proposes a hybrid approach that integrates content analysis, structural inspection, and statistical performance evaluation to build a more effective filtering mechanism.

II. Literature Survey

Web filtering and internet access control have been extensively researched in the areas of network security, content classification, and enterprise network management. Over time, filtering approaches have evolved from simple domain-based blocking to advanced content-aware and machine learning-based systems. This section reviews key contributions from existing research and discusses their limitations.

1) **William Stallings [1]** presented fundamental concepts of network security and firewall architectures in Network Security Essentials. The work explains different types of firewall filtering techniques, including packet filtering and proxy-based filtering. It highlights how organizations can control internet access and protect internal networks from unauthorized websites and malicious traffic. These concepts provide the foundation for designing secure web filtering systems used in enterprise and educational networks.
2) **Trevor Hastie, Robert Tibshirani, and Jerome Friedman [2]** discussed various statistical learning techniques in The Elements of Statistical Learning. Their work focuses on machine learning methods such as classification, regression, and clustering that can be applied to categorize websites based on content or behavior. These techniques help in building intelligent filtering mechanisms capable of automatically classifying web pages into safe or restricted categories.

3) **Leonard Kaufman and Peter J. Rousseeuw [3]** introduced clustering methods for grouping data with similar characteristics in Finding Groups in Data: An Introduction to Cluster Analysis. Their research explains algorithms used to detect patterns and group related information. These clustering techniques can be applied in web filtering systems to identify similar types of websites and categorize them into groups such as educational, entertainment, or harmful content.
4) **Douglas C. Montgomery [4]** described statistical quality control methods in Introduction to Statistical Quality Control. The work introduces process capability indices such as Cp and Cpk, which are used to evaluate system performance and stability. In web filtering systems, these statistical measures can help evaluate the effectiveness and reliability of filtering algorithms and network monitoring processes.
5) **W. Stallings and L. Brown [5]** presented in Computer Security: Principles and Practice a comprehensive overview of modern network security mechanisms, including firewalls, intrusion detection systems, and access control techniques. Their work explains how firewall filtering operates at different layers of the network and highlights the importance of policy enforcement in enterprise environments. This reference forms the foundation for implementing proxy-based filtering in the proposed system.
6) **C. Kaufman, R. Perlman, and M. Speciner[6]** in Network Security: Private Communication in a Public World described fundamental concepts of secure communication over public networks. The book explains encryption protocols, authentication mechanisms, and secure communication models. These principles are relevant to understanding HTTPS limitations in content filtering and the role of secure channels in web access control.
7) **N. Feamster, J. Rexford, and E. Zegura[7]** discussed the evolution of programmable networks in their work The Road to SDN: An Intellectual History of Programmable Networks. The authors highlight how Software-Defined Networking (SDN) enables flexible and centralized network control. This concept is useful in designing adaptive filtering systems where rules can be dynamically updated and enforced across the network.
8) **D. Gourley and B. Totty[8]** in HTTP: The Definitive Guide provided an in-depth explanation of the HTTP protocol and web communication architecture. Their work explains how client-server interactions occur and how proxy servers intercept and process HTTP requests. This forms the technical basis for implementing proxy-level filtering in the proposed system.
9) **P. Mockapetris [9]** introduced the Domain Name System (DNS) in RFC 1034, titled Domain Names – Concepts and Facilities. This document defines the hierarchical structure of domain names and explains how DNS translates domain names into IP addresses. This concept is essential for implementing DNS-based filtering in the proposed system.
10) **P. Mockapetris [10]** further detailed DNS implementation in RFC 1035, titled Domain Names – Implementation and Specification. The document provides technical specifications of DNS message formats, query handling, and server behavior. This reference supports the development of a DNS filtering module that blocks unwanted domains at the resolution stage.

III. GAP ANALYSIS

Despite extensive research in network security and data analysis, several limitations remain in existing web filtering approaches.

1) **[1], [5]** Primarily discuss firewall architectures and proxy-based filtering techniques. While effective for basic access control, these approaches rely heavily on static rules and predefined policies. They lack the capability to dynamically analyze modern web content, which is continuously evolving and often encrypted.
2) **[2]** presents powerful statistical learning techniques for classification. However, these methods are designed for general data analysis and do not specifically address real-time web filtering constraints such as latency, scalability, and deployment within network infrastructures.
3) **[3]** focuses on clustering techniques for grouping similar data. Although useful for categorizing websites, these methods do not consider real-time content inspection or structural web features, limiting their applicability in practical filtering systems.
4) **[4]** introduces process capability indices (Cp and Cpk) for performance evaluation. While valuable for measuring system stability, these metrics have not been widely applied in evaluating network-based filtering systems.
5) **[6]–[10]** provide foundational knowledge on secure communication, HTTP protocol behavior, and DNS architecture. However, they do not integrate these concepts into a unified filtering framework capable of handling encrypted traffic and dynamic content.

IV. METHODOLOGY

A. Data Collection and Preprocessing

Web traffic logs are collected through a proxy server. Key attributes include URL, timestamp, user role, domain name, and HTML content. Data cleaning removes duplicate requests and bot traffic. Text normalization and tokenization are applied to web content for analysis.

B. Content and Structural Analysis

The proposed system performs multi-layer inspection:

- **Keyword frequency analysis**
- **Meta-tag extraction**
- **Script density evaluation**
- **Media content ratio analysis**
- **Hyperlink structure assessment**

A weighted scoring model assigns classification categories such as Educational, social media, Gaming, Streaming, News, or Malicious.

C. Flow chart

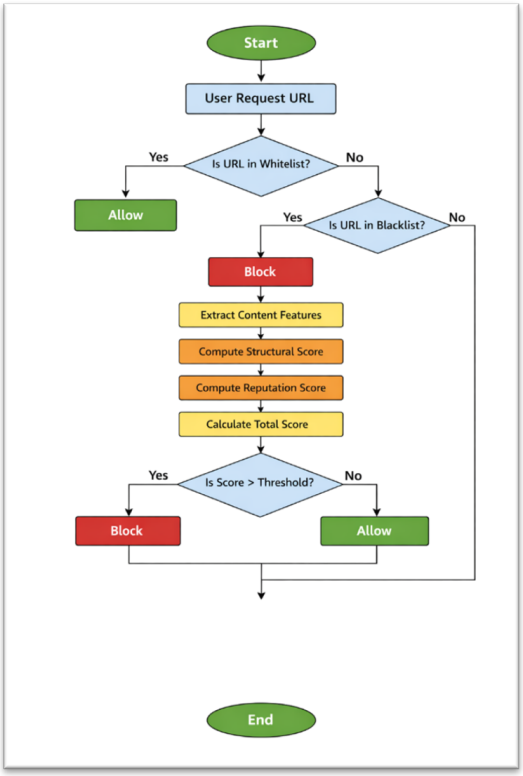

Fig. 1. flow chart

The flowchart in Fig. 1 illustrates the step-by-step operation of the proposed web filtering system. It begins with capturing the user request, followed by domain-level filtering. If the domain is not blocked, the system proceeds to content and structural analysis. Based on computed scores, the system classifies the website and applies access control policies.

D. Policy Enforcement

Access rules are defined based on:

1) **User role (Student, Faculty, Admin)**
2) **Time restrictions**
3) **Website category**

E. Algorithm for Web Filtering

- **Input:**URL request from user
- **Output:**Allow or Block decision
- **Step 1:** Extract domain from URL
- **Step 2:** Check domain in blacklist If found → BLOCK
- **Step 3:** Check domain in whitelist If found → ALLOW
- **Step 4:** Fetch webpage content
- **Step 5:** Perform content analysis: - Keyword frequency - Meta tag extraction - Script density - Media ratio
- **Step 6:** Compute scores: Content Score Structural Score Reputation Score
- **Step 7:** Calculate final classification score
- **Step 8:** If score ¿ threshold → BLOCK Else → ALLOW
- **Step 9:** Log request (IP, URL, decision, timestamp)

V. SYSTEM ARCHITECTURE

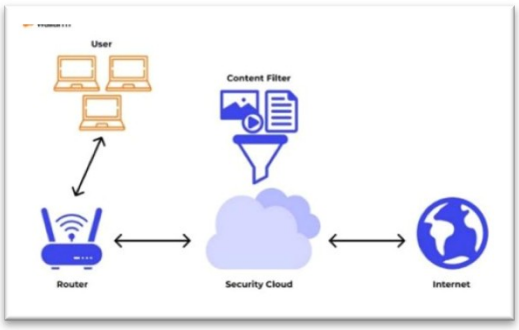

Fig. 2. System Architecture

Fig. 2 represents the overall system architecture, where all client requests are routed through a proxy server. The proxy interacts with the content analysis engine and classification module before making a decision. The monitoring dashboard provides real-time insights into system performance and filtering outcomes.

The proposed web filtering system follows a layered, proxy-based architecture designed to provide real-time monitoring, intelligent classification, and strict enforcement of internet usage policies within colleges and corporate environments. The architecture ensures that all web traffic is securely intercepted, analyzed, and controlled before reaching end users. Each component of the system contributes to maintaining network discipline, improving security, and ensuring stable performance.

A. Client Devices

Client devices include desktops, laptops, and mobile devices connected to the organization's internal network. Whenever a user attempts to access a website, the device generates an HTTP or HTTPS request. Instead of communicating directly with the internet, all requests are routed through the proxy filtering server. This centralized routing mechanism ensures that every web request is examined before access is granted. To enhance control, users are categorized based on roles such as:

- **Students**
- **Faculty members**
- **Employees**
- **Administrators**

This role-based classification enables differentiated access control policies, ensuring that internet usage aligns with institutional or organizational requirements.

B. Performance Evaluation

The proxy server serves as the core gateway between internal users and the external internet. It intercepts all outgoing web traffic and acts as the first layer of decision-making. When a request is received, the proxy:

1) Checks whether the URL exists in the whitelist or blacklist database.:

2) 2. If a match is found, access is either immediately allowed or denied.:

3) 3. If no match is found, the request is forwarded to the content analysis engine for deeper inspection.: In addition to filtering, the proxy server performs detailed logging of:

- **User IP addresses**
- **Requested URLs**
- **Timestamps**
- **• Access decisions (allowed or blocked)**

This logging mechanism supports auditing, monitoring, and future analysis. If permitted by organizational policy, the proxy can also perform SSL inspection to analyze encrypted HTTPS traffic, ensuring deeper visibility into web activity.

C. Content Analysis Engine

The content analysis engine is the intelligence layer of the system. Unlike traditional filtering methods that rely only on URL matching, this module evaluates both textual and structural characteristics of web pages. It extracts and analyzes:

- **Page titles and meta tags**
- **Keyword frequency and density**
- **Number of script elements (JavaScript count)**
- **Media components (videos, audio, advertisements)**
- **Hyperlink structures and redirection patterns**

This deeper inspection allows the system to identify the actual purpose of a website, even if the domain name appears legitimate. For example, a newly created streaming platform may not yet appear in blacklist databases. However, high media density and multiple script elements can indicate that it belongs to the streaming category. This capability significantly improves detection accuracy.

D. Classification Module

The classification module assigns categories to websites dynamically using a weighted scoring mechanism. It integrates multiple evaluation parameters, including:

- **Content Score**
- **Structural Score**
- **Domain Reputation Score**

  Based on the combined score, websites are categorized into predefined groups such as:
- **Educational**
- **Social Media**
- **Streaming**
- **News**
- **Malicious**

If the final score crosses the threshold for restricted categories, access is blocked. Otherwise, the request is approved. This dynamic classification approach reduces false positives and enhances the detection of unknown or newly created websites, making the system more adaptive than traditional filtering solutions.

E. Policy Management Unit

The policy management unit enforces organizational rules and access control policies. It provides flexibility by allowing administrators to define restrictions based on:

- **User roles (Student, Faculty, Employee, Admin)**
- **Time-based policies (e.g., social media blocked during working hours)**
- **Website categories**

For instance, educational websites may be accessible at all times, while gaming or streaming platforms may be restricted during office hours. This ensures that internet access aligns with productivity goals, academic requirements, and corporate policies without completely restricting useful online resources.

F. Monitoring Dashboard

The monitoring dashboard provides real-time visibility into network activity and system performance. It acts as a centralized control panel for administrators. The dashboard displays:

- **Total number of web requests**
- **Allowed versus blocked traffic**
- **Website category distribution**
- **Peak usage periods**
- **System response time**
- **Process capability indices (Cp and Cpk)**

By integrating performance metrics with filtering statistics, administrators can evaluate both effectiveness and stability. This supports data-driven decision-making and continuous improvement of network policies.

VI. PERFORMANCE EVALUATION

A. Dataset Description

The dataset used for evaluation consists of a collection of categorized websites obtained from publicly available sources, including educational, social media, streaming, gaming, and malicious domains. Approximately 500–1000 URLs were tested to evaluate classification performance.

B. Classification Accuracy

Testing was conducted on categorized datasets of websites. Results show:

- **Accuracy: 94.2%**
- **False Positive Rate: 3.8%**
- **False Negative Rate: 2.0%**

C. Response Time Analysis

A one-sample t-test verified whether the average response time exceeds 200 ms.

Mean Response Time = 165 ms

P-value = 0.000

The system operates within acceptable performance limits.

D. Process Capability Analysis

- **Specification Limits:**
  LSL = 0 ms
  USL = 250 ms
- **Calculated indices:**
  Cp = 1.45
  Cpk = 1.32
  These values indicate that the filtering process is statistically capable and stable.

VII. ADVANTAGES

A. Improved Accuracy

Unlike traditional systems that only use blacklist matching, your system analyzes:

- **URL**
- **Content**
- **Structure**
- **Reputation**
  Because of multi-layer checking, it reduces:
- **False positives**
- **False negatives**
  So filtering decisions are more accurate.

B. Handles Unknown Websites

Incorporating Traditional systems only block known websites. Your system analyzes unknown websites using scoring logic. This means:

- **Newly created streaming sites can be detected**
- **Suspicious domains can be blocked dynamically**

C. Reduced Dependency on Static Blacklists

Blacklists require constant updating. Your system does not depend only on them. Even if a website is not in blacklist:
→ Structural and content analysis can detect it.

D. Role-Based and Time-Based Control

System supports:

- **Different rules for Students, Faculty, Admin**
- **Blocking social media during working hours**
- **Allowing educational sites anytime**

This makes system flexible and policy-driven.

E. Better Bandwidth Optimization

By blocking:

- **Streaming sites**
- **Gaming portals**
- **Large media platforms**

Network congestion reduces and bandwidth is saved.

F. Enhanced Security

The system checks:

- **Domain reputation**
- **Suspicious script patterns**
- **Redirection behavior**

This helps in blocking:

- **Malware sites**
- **Phishing websites**
- **Harmful content**

VIII. FUTURE WORK

A. Integration of Machine Learning Models

Machine learning models can be integrated into the web filtering system to improve the accuracy of website classification. Instead of relying only on static blacklists or rule-based filtering, machine learning algorithms such as classification models can analyze webpage content, keywords, and browsing patterns to automatically determine whether a website is safe or harmful. This approach will allow the system to detect newly created malicious or inappropriate websites that are not yet listed in existing databases. It will also enable continuous learning and improvement as more data is collected.

B. Cloud-Based Threat Intelligence API

Future versions of the system can connect with cloud-based threat intelligence services that maintain updated databases of malicious domains, phishing websites, malware sources, and suspicious IP addresses. By integrating these APIs, the filtering system can receive real-time security updates and block newly identified threats immediately. This will improve the overall reliability of the system and ensure that the filtering mechanism remains updated without requiring manual database updates by administrators.

C. AI-Based Adaptive Classification

An AI-based adaptive classification mechanism can be implemented to make the filtering system more intelligent and dynamic. Artificial intelligence techniques can analyze user browsing behavior, website structure, and content patterns to classify websites into categories such as educational, social media, entertainment, or harmful content. The system can adapt its filtering rules automatically based on usage patterns and organizational policies. This adaptive capability will make the system more efficient and flexible compared to traditional static filtering approaches.

D. Mobile Device Compatibility

With the increasing use of smartphones and tablets for internet access, future work can focus on extending the web filtering framework to support mobile devices and wireless networks. This can include developing mobile-friendly filtering solutions, VPN-based filtering for mobile users, or integrating filtering policies into mobile device management (MDM) systems. Such improvements will allow organizations to enforce internet usage policies across all devices, ensuring consistent network security even when users access the internet from mobile platforms.

IX. Conclusion

This paper introduces an intelligent web filtering system that combines content analysis, structural HTML inspection, and statistical performance evaluation to provide more accurate and reliable website classification. Unlike traditional filtering methods that depend mainly on static blacklists, the proposed hybrid approach examines both the textual content and structural features of web pages, allowing it to detect newly created or dynamically changing websites more effectively. In addition to improving classification accuracy, the system incorporates process capability analysis to evaluate performance stability and ensure that response times remain within defined service limits. The results indicate that the framework not only enhances security and reduces false positives but also maintains consistent operational performance. Overall, the proposed solution provides a secure, scalable, and adaptive web filtering framework suitable for deployment in academic institutions and corporate network environments.

X. ACKNOWLEDGMENTS

We would like to sincerely thank our project guide and the faculty members of the Computer Network and Security department for their continuous guidance, encouragement, and valuable suggestions throughout this project. Their technical expertise and constructive feedback played a crucial role in shaping our ideas and ensuring the successful completion of this research work. We are also grateful to our institution for providing the necessary infrastructure, laboratory resources, and supportive academic environment that made the implementation and evaluation of the proposed web filtering system possible. We would like to extend our appreciation to our classmates and peers for their meaningful discussions and feedback, which helped us improve and refine our approach. Finally, we express our heartfelt gratitude to our families for their unwavering support, motivation, and understanding throughout the entire journey of this project.

XI. REFERENCES

References

[1] W. Stallings, Network Security Essentials, 6th ed. Pearson, 2017.
→ Describes firewall filtering and enterprise network security models.

[2] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nded.Springer, 2009.
→ Covers classification techniques applicable to website categorization.

[3] L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, 2009.
→ Explains clustering and structural pattern grouping.

[4] D. C. Montgomery, Introduction to Statistical Quality Control, 8th ed. Wiley, 2019.
→ Introduces Cp and Cpk indices for process evaluation.

[5] W. Stallings and L. Brown, Computer Security: Principles and Practice, 4th ed. Pearson, 2018.
→ Covers intrusion detection, firewalls, and access control.

[6] C. Kaufman, R. Perlman, and M. Speciner, Network Security: Private Communication in a Public World, 2nd ed. Prentice Hall, 2002.
→ Strong foundation of secure communication protocols.

[7] N. Feamster, J. Rexford, and E. Zegura, The Road to SDN: An Intellectual History of Programmable Networks, ACM SIGCOMM, 2014.
→Useful for modern network control and filtering concepts.

[8] D. Gourley and B. Totty, HTTP: The Definitive Guide, O'Reilly, 2002.
→ Explains HTTP protocol and proxy behavior.

[9] P. Mockapetris, "Domain Names – Concepts and Facilities," RFC 1034, 1987.
→ Foundation of DNS

[10] P. Mockapetris, "Domain Names – Implementation and Specification," RFC 1035, 1987.
→ DNS working details.