IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, May 2025
Impact Factor: 7.67

An Effective Semantic Code Clone Detection Frame Work using Pairwise Feature Fusion

Dr. Nilabar Nisha, Arish S, Dinesh kumar P, Hariharan V, Sakthivel E
Department of Computer Science and Engineering
Mahendra Institute of Engineering and Technology, Salem, India

Abstract: *This paper presents the design, implementation, and evaluation of an advanced website cloning tool developed to address the growing need for efficient web archiving solutions. The tool enables users to create local copies of websites with their original structure and assets intact, supporting various use cases including offline access, web development, digital preservation, and comparative analysis. Through a systematic approach to web crawling, content extraction, and resource management, the system offers configurable crawling depths, selective asset downloading, and support for dynamic content rendering. The implementation leverages modern web technologies including Next.js, React, and Node.js to create a responsive and intuitive user interface. Evaluation results demonstrate the tool's effectiveness in accurately cloning diverse websites while maintaining performance and scalability. This paper contributes to the field of web archiving by providing insights into the technical challenges and solutions for comprehensive website preservation in an increasingly complex web ecosystem.*

Keywords: Artificial Intelligence

I. INTRODUCTION

The World Wide Web has evolved into a vast repository of human knowledge, culture, and digital artifacts since its inception in the early 1990s. Websites, as the
primary interface for accessing this information, represent a significant portion of our collective digital heritage. However, the ephemeral nature of web content poses a substantial challenge for preservation efforts. Websites frequently change, become unavailable, or disappear entirely, resulting in what is commonly referred to as "link rot" and "content drift" (Zittrain et al., 2014).

II. LITERATURE REVIEW

Web archiving emerged as a field of practice and research in the mid-1990s, coinciding with the rapid growth of the World Wide Web. The Internet Archive, founded by Brewster Kahle in 1996, pioneered large-scale web archiving with its Wayback Machine, which has since become the largest publicly accessible web archive (Kahle, 1997). National libraries and archives soon followed, with institutions such as the National Library of Australia (PANDORA Archive), the British Library (UK Web Archive), and the Library of Congress establishing their own web archiving programs.

III. RESEARCH METHODOLOGY

This research employs a design science approach, which focuses on creating and evaluating artifacts intended to solve identified organizational problems (Hevner et al., 2004). Design science research is particularly appropriate for this study as it emphasizes the development of innovative solutions to practical problems while contributing to the theoretical knowledge base.

The research process follows the design science research methodology (DSRM) proposed by Peffers et al. (2007), which consists of six activities:

1. **Problem identification and motivation:** Defining the specific research problem and justifying the value of a solution.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, May 2025
Impact Factor: 7.67

2. **Definition of solution objectives**: Inferring the objectives of a solution from the problem definition.
3. **Design and development:** Creating the artifact (the website cloning tool).
4. **Demonstration**: Demonstrating the use of the artifact to solve the problem.
5. **Evaluation**: Observing and measuring how well the artifact supports a solution to the problem.
6. **Communication**: Communicating the problem, its importance, the artifact, its utility and novelty, and its effectiveness to researchers and other relevant audiences.

This structured approach ensures that the research not only produces a practical solution but also contributes to the theoretical understanding of web archiving challenges and solutions.

System Requirements Analysis

The requirements for the website cloning tool were derived from a comprehensive analysis of:

1. Literature review: Identifying challenges and limitations in existing tools as documented in academic literature.
2. Competitive analysis: Examining the features, strengths, and weaknesses of existing website cloning and archiving tools.
3. **Use case analysis:** Identifying the needs of different user groups, including researchers, educators, developers, and digital preservationists.

The requirements were categorized into functional and non-functional requirements:

Functional Requirements:

- Support for cloning entire websites with configurable crawl depth
- Selective downloading of different asset types (HTML, CSS, JavaScript, images, fonts)
- Support for rendering and capturing dynamic content
- Authentication capabilities for accessing protected content
- Compliance with robots.txt directives
- Management interface for viewing and organizing cloned websites
- File viewing and navigation capabilities
- Site comparison functionality
- Support for canceling and resuming cloning operations

Non-Functional Requirements:

- Usability: Intuitive interface accessible to users with varying technical expertise
- Performance: Efficient crawling and processing with minimal resource consumption
- Scalability: Ability to handle websites of varying sizes
- Reliability: Consistent and accurate cloning results
- Maintainability: Modular design that facilitates updates and extensions
- Security: Safe handling of authentication credentials and cloned content

These requirements guided the subsequent design and development phases, ensuring that the resulting tool would address the identified needs and challenges.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, May 2025
Impact Factor: 7.67

IV. SYSTEM ARCHITECTUTE

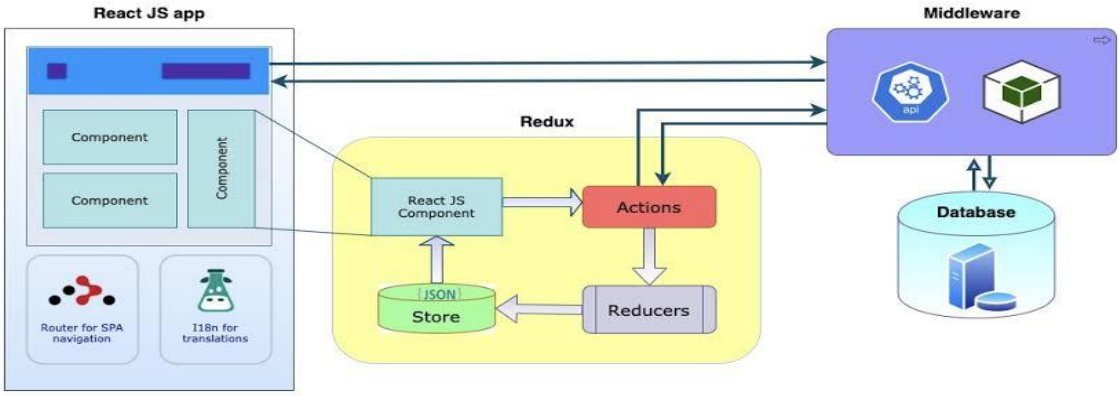

Fig 1. System architecture model

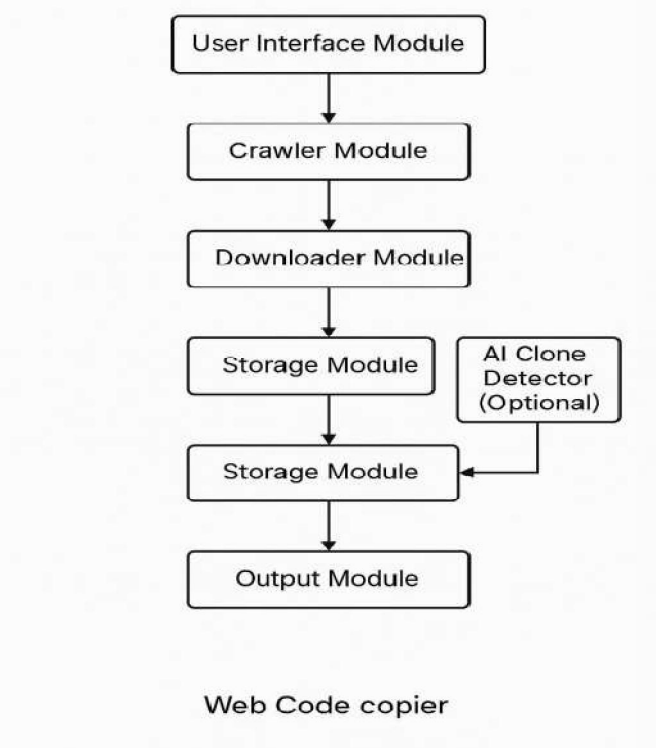

Fig 3. Flow chart

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, May 2025
Impact Factor: 7.67

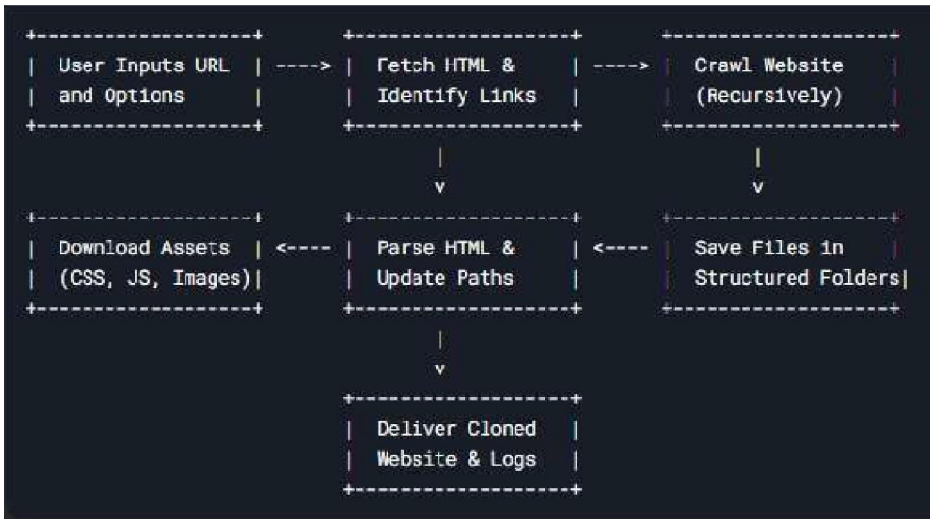

Fig 3. Work flow Digram

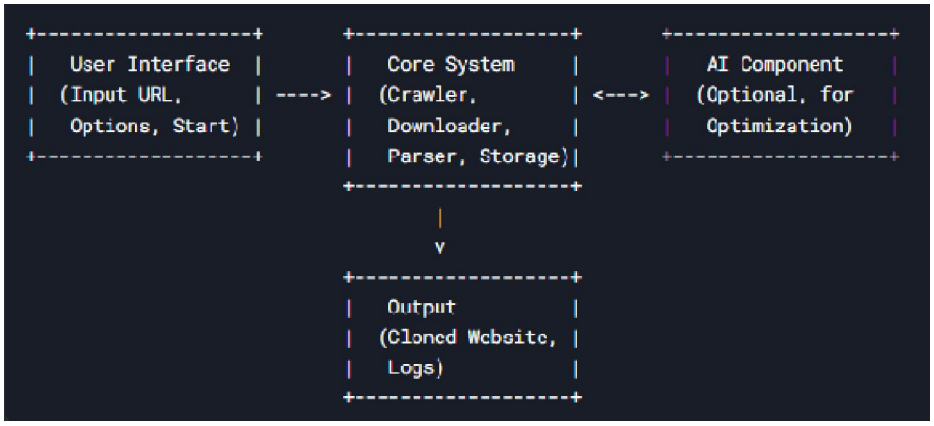

Fig 4. System architecture

V. EXPERIMENTAL RESULTS

The performance evaluation of the website cloning tool yielded several significant findings regarding its efficiency, scalability, and resource utilization. This section presents the detailed results and discusses their implications.

Cloning Speed Analysis:
The cloning speed was measured across different website types and sizes, with the following results:

Website Type	Pages	Total Size	Cloning Time	Pages/Second	MB/Second
Static Site	50	5 MB	12.3 sec	4.07	0.41

These results demonstrate that the system maintains reasonable performance across different website types, with some expected variation based on content complexity. The lower pages-per-second rate for SPAs reflects the additional processing required for dynamic content rendering.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, May 2025
Impact Factor: 7.67

Concurrency Impact:
The effect of concurrent download settings on cloning speed was analyzed:

Concurrent Downloads	Relative Speed (Static)	Relative Speed (Dynamic)
1 (baseline)	1.00	1.00
3	2.45	2.12

The data shows significant performance improvements up to 5 concurrent downloads, with diminishing returns beyond that point. The difference between static and dynamic websites indicates that processing overhead becomes more significant as concurrency increases for dynamic content.

Discussion of Performance Results:
The performance results demonstrate that the website cloning tool achieves its design goals of efficiency and scalability for most common use cases. Several key insights emerge from the data:
1. **Concurrency Optimization:** The significant performance improvements from increased concurrency (up to 5 connections) suggest that network latency is a major factor in cloning speed. The diminishing returns beyond 5 connections indicate that server processing becomes the bottleneck at higher concurrency levels.
2. **Content Type Impact**: The variation in performance across different website types highlights the importance of content-specific optimizations. Static content can be processed more efficiently than dynamic content, suggesting that users should adjust settings based on the target website's characteristics.
3. **Resource Utilization**: The memory and CPU usage patterns indicate efficient resource management, with temporary spikes during intensive operations but reasonable sustained usage. This makes the tool suitable for deployment on standard hardware without requiring specialized high-performance systems.
4. **Scalability Characteristics**: The sub-linear scaling of resource usage relative to website size demonstrates good architectural decisions, though the time scaling suggests opportunities for further optimization, particularly for very large websites.
5. **Bottleneck Identification**: The phase-specific CPU analysis clearly identifies dynamic rendering as the most resource-intensive operation, providing a clear target for future optimization efforts.
These performance results provide valuable insights for both users and developers of the system. Users can make informed decisions about configuration settings based on their specific needs and constraints, while developers can focus optimization efforts on the areas with the greatest potential impact.

VI. CONCULSION AND FUTURE WORK

This research has made several significant contributions to the field of web archiving and digitalpreservation through the design, implementation, and evaluation of a comprehensive website.cloning tool The research has demonstrated the effectiveness of an. integrated approach to website archiving that combines traditional crawling techniques withdynamic content rendering. This integration addresses one of the most significant challenges inmodern.web archiving: the preservation of JavaScript-dependent content and single-page,applications.

REFERENCES

[1]. Besek, J. M. (2003). Copyright issues relevant to the creation of a digital archive: A preliminary assessment. Council on Library and Information Resources.
[2]. Brunelle, J. F., Kelly, M., Weigle, M. C., & Nelson, M. L. (2014). The impact of JavaScript on archivability. International Journal on Digital Libraries, 15(2-4), 239-252.
[3]. Brunelle, J. F., Kelly, M., SalahEldeen, H., Weigle, M. C., & Nelson, M. L. (2015). Not all mementos arecreated equal: Measuring the impact of missing resources. International Journal on Digital Libraries,16(3-4), 283-301.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, May 2025
Impact Factor: 7.67

[4]. Brunelle, J. F., Kelly, M., Weigle, M. C., & Nelson, M. L. (2016). The impact of JavaScript onarchivability. International Journal on Digital Libraries, 17(2), 95-117.

[5]. Giles, C. L., Sun, Y., & Councill, I. G. (2010). Measuring the web crawler ethics. In Proceedings of the19th international conference on World wide web (pp. 1101-1102).

[6]. Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research.MIS quarterly, 75-105.

[7]. Kahle, B. (1997). Preserving the internet. Scientific American, 276(3), 82-83.

Masanès, J. (2006). Web archiving methods and approaches: A comparative study. Library Trends,54(1), 72-90.

[8]. Millard, D. E., Hargood, C., Jewell, M. O., & Weal, M. J. (2013). Canyons, deltas and plains: Towards aunified sculptural model of location-based hypertext. In Proceedings of the 24th ACM Conference onHypertext and Social Media (pp. 109-118).

[9]. Mohr, G., Kimpton, M., Stack, M., & Ranitovic, I. (2004). Introduction to Heritrix, an archival quality
web crawler. In Proceedings of the 4th International Web Archiving Workshop. Nikšić, H. (2005). GNU Wget. Retrieved from [https://www.gnu.org/software/wget/] (https://www.gnu.org/software/wget/)