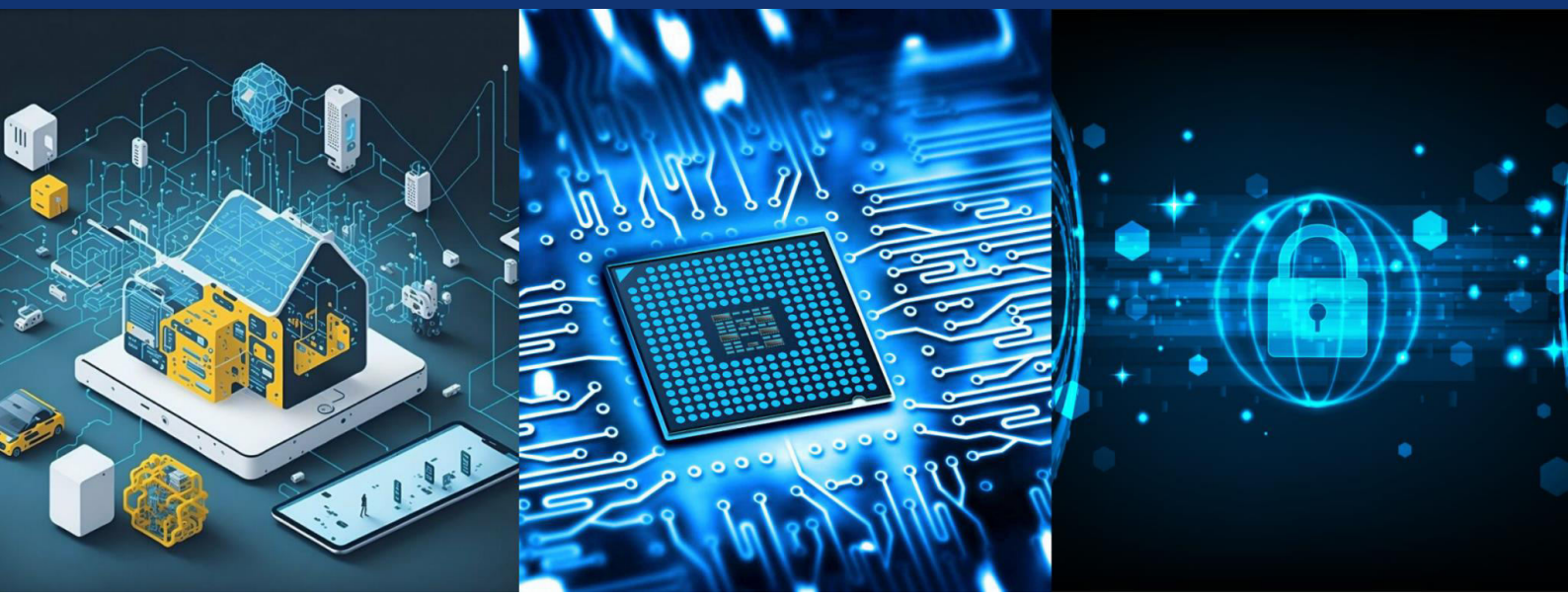
IJIRCCE

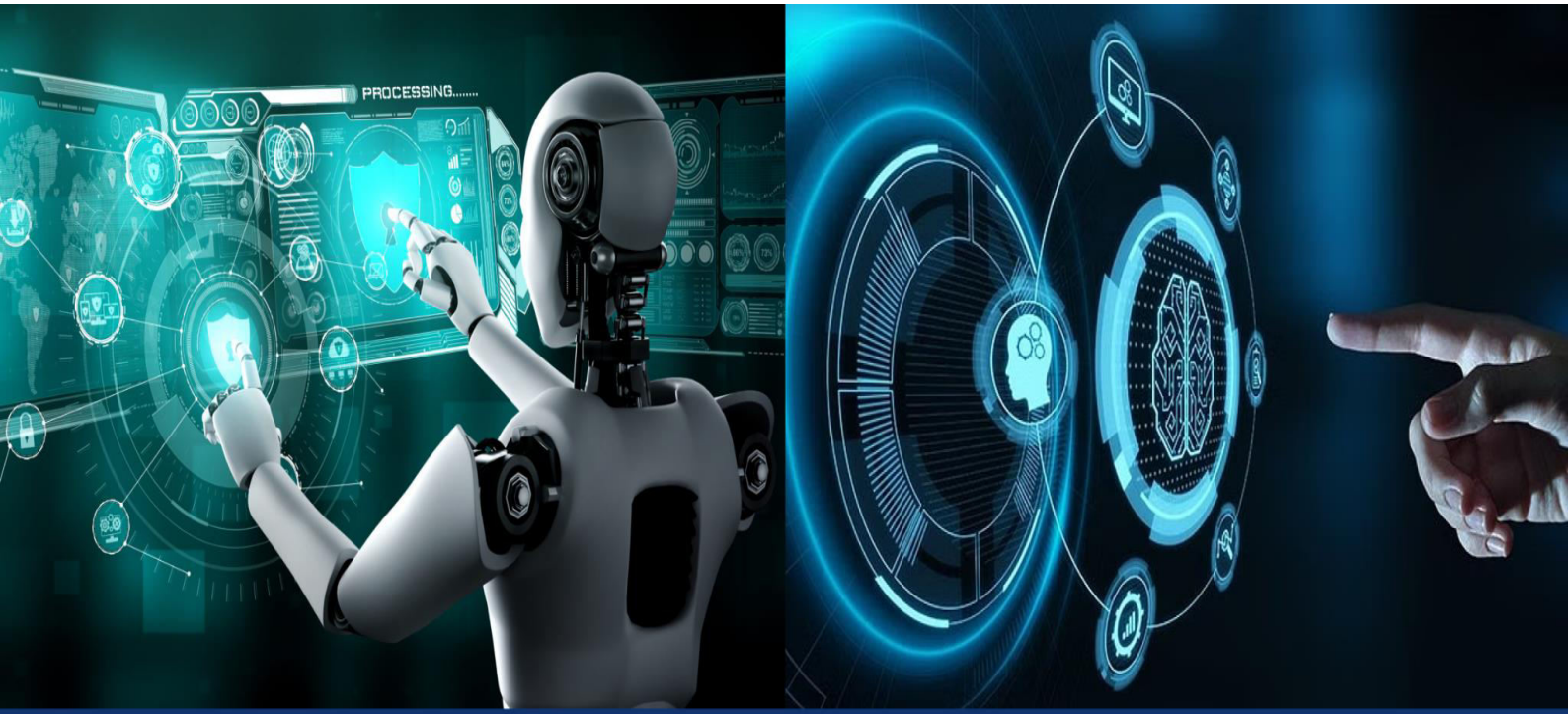
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.771

Volume 13, Issue 4, April 2025

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI10.15680/IJIRCCE.2025.1304109

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Advanced Streamlined Crawling

Mrs. Preethi Reddy[1], V A. Revanth[2], K. Rakesh[3], K. Ram Reddy[4], R. Sai Karthik[5]

Guide, Department of Computer Science, Malla Reddy University, Hyderabad, Telangana, India[1]

Department of Computer Science, Malla Reddy University, Hyderabad, Telangana, India[2-5]

ABSTRACT: The World Wide Web (WWW) is a vast, dynamic, and constantly evolving resource of information, with a significant portion of its content residing in the deep web—sections of the internet not indexed by traditional search engines. Extracting meaningful data from the deep web presents numerous challenges due to its dynamic nature, hidden directories, and the large volume of resources required for data collection. Traditional crawlers often struggle to effectively navigate these obstacles, making it difficult to gather relevant information efficiently. To address these challenges, a two-stage web crawler has been developed. The primary goal of this project is to create an efficient and effective solution for gathering high-quality data from the deep web.By utilizing a well-organized process to gather and scrape data, the two-stage crawler offers a valuable tool for applications in academic research, market analysis, and other data-driven fields. It demonstrates how targeted, efficient crawling of the WWW can overcome the limitations of traditional search engines, unlocking valuable, otherwise hidden information from less-visible parts of the internet.

I.INTRODUCTION

A web crawler is a system for the bulk downloading of web pages. Web crawlers are used for a variety of purposes. Most prominently, they are one of the main components of web search engines, systems that assemble a corpus of web pages, index them, and allow users to issue queries against the index and find the web pages that match the queries.
It is defined as a program or software which traverses the Web and downloads web documents in a methodical, automated manner. Based on the type of knowledge, web crawler is usually divided in three types of crawling techniques: General Purpose Crawling, Focused crawling and Distributed Crawling..

The deep web is the portion of the internet that is not indexed by traditional search engines. It has databases, private sites, and other resources that are hidden behind forms, passwords, or dynamic content. Since its content is not directly accessible through normal web crawlers, specialized crawlers are used to locate and extract data from those hidden databases, enabling users to search for and access information that would otherwise be difficult to find.

The significance of web crawlers extends beyond basic data collection; they address the inherent challenges and complexities of organizing and indexing vast amounts of online information. Traditional methods of manually gathering and updating data can be time-consuming and prone to error, often leading to inefficiencies and inconsistencies. Web crawlers, particularly two-stage crawlers, offer a more structured, automated, and scalable solution to these problems. In a two-stage process, the discovery of URLs is separated from the content retrieval, optimizing resources and ensuring faster, more efficient data collection. This method streamlines web crawling, reducing errors and enhancing performance when handling dynamic and static websites.

II.LITERATURE REVIEW

In two-stage crawling, the first phase typically involves identifying and fetching a large set of initial URLs, often based on heuristics or predefined search criteria. This process aims to cover a broad range of potential sources. The second phase, however, refines the crawling process by analyzing and selecting more specific or deeper URLs based on factors like content relevance, site structure, or link structure. [1] proposed a model for two-stage crawlers that emphasized the importance of controlling the crawling depth. Their work underscored the need to prioritize pages likely to contain more relevant or high-quality content, which in turn improves the efficiency of the crawling process without sacrificing the quality of collected data. Similarly, [4] highlighted how two-stage crawlers are crucial in search engines for relevance filtering, particularly when dealing with large-scale document repositories. Their research has had a lasting impact on the design of modern web crawlers and search engine architectures. More recently, [3] discussed how deep learning and reinforcement learning could enhance the second stage of crawling, improving the prediction of relevant pages and optimizing resource usage.[5] also explored the use of two-stage crawlers in the e-commerce sector,

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI10.15680/IJIRCCE.2025.1304109

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

specifically for gathering product information across multiple online retailers. They demonstrated that focusing on more relevant product pages in the second phase significantly enhances data extraction. Lastly, [2] proposed a strategy that optimizes URL selection in the second stage through machine learning methods, using historical data to predict which pages are more likely to be valuable based on content type and structure, thus minimizing unnecessary crawling.

III. METHODOLOGY

The objective is to provide a comprehensive understanding of how web crawlers are optimized to locate relevant sites and conduct focused crawling for specific data extraction, such as searchable forms.

Initial Crawling (Broad Crawling)

The initial crawling phase is foundational for broad web exploration. During this phase, the objective is to explore a wide variety of websites to gather as many URLs as possible without delving into content analysis. This phase typically follows a process involving the generation of seed URLs, web page fetching, link extraction, and URL storage.

URL Seed Generation: A significant body of research [1] discusses the process of starting with a seed set of URLs obtained from various sources, such as predefined URL lists, search engines, web directories, and user-provided URLs. The aim of this step is to gather diverse URLs that may lead to relevant pages for further exploration.

Web Page Fetching: Once the seed URLs are identified, the crawler fetches web pages using HTTP requests. This step involves determining the crawl depth, which is often large during the initial crawling to maximize the collection of links [3]. The crawl depth is managed to balance the efficiency of the crawler with the relevance of collected data.
Politeness Policy: According to [4] politeness policies are critical for maintaining crawler efficiency and respecting website limitations. This includes rate-limiting requests and adhering to robots.txt files to prevent overloading websites and maintain ethical crawling practices.

Link Extraction and Filtering: The extracted hyperlinks are filtered based on predefined criteria, such as excluding non-HTML links like PDFs or images and ensuring the links are valid [5] This step ensures that only relevant and useful links are retained for further processing.

URL Storage: The valid URLs are stored in a crawl frontier (URL queue). The queue is often prioritized using various strategies such as FIFO (First-In-First-Out) or priority queues, as emphasized in the work of [2] which focuses on optimizing the crawling process for efficient data collection.

Crawl Frontier Management: The URL queue is managed through depth limitation and duplicate removal techniques. Limiting the crawl depth ensures that irrelevant content is not fetched, and duplicate URLs are removed to optimize resource usage [4].

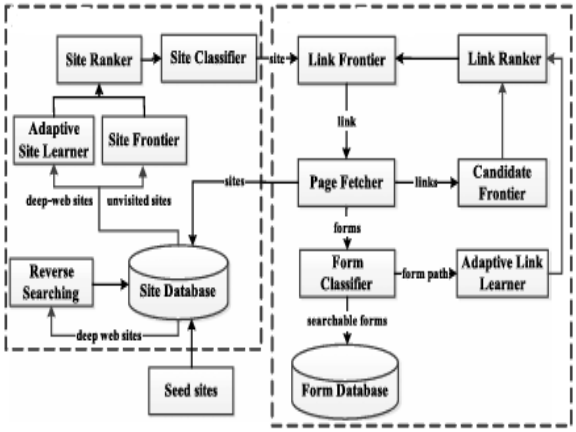


Fig1 -Two stage Architecture

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Refined Crawling (Focused Crawling)

Once a broad set of URLs has been collected during Stage 1, the focus shifts to refining the crawling process. The second stage analyzes the content of the web pages and prioritizes URLs based on their relevance to the desired topic or objective.

Relevance Analysis of URLs: The relevance of URLs is evaluated through various content-based filtering techniques. These include:

Textual Content Analysis: By analyzing the text of a webpage, crawlers determine if the page content is relevant to predefined keywords or topics [3].

Metadata Extraction: Crawlers extract metadata such as titles, descriptions, and keywords, comparing it with predefined themes to assess relevance [2].

Machine Learning: In recent years, machine learning techniques have been applied to classify URLs based on patterns from previously crawled pages. This adaptive approach is discussed by **[2]**, who highlight the effectiveness of machine learning in predicting the relevance of pages.

Dynamic URL Prioritization: Once URLs are analyzed, a relevance score is applied to prioritize the most relevant pages. Scoring mechanisms typically take into account:

- The frequency of keywords in the content,
- Semantic similarity to the desired topic,
- The link structure and trustworthiness of the linking page [5].

URLs with higher relevance scores are prioritized for crawling, ensuring that resources are focused on the most valuable content.

Focused Crawling: In this stage, the crawler selectively requests pages based on the relevance scores assigned in the previous step. By focusing on high-priority URLs, the crawler minimizes the fetching of irrelevant content [1]. This approach is crucial for efficient resource usage and for gathering high-quality data.

Advanced techniques are employed to further refine the crawling process:

URL Clustering: Clustering algorithms, such as k-means, can group URLs based on similarity. This helps the crawler avoid redundant content and identify new domains with similar topics [3].

Adaptive Crawling: Feedback from the crawling process allows for dynamic adjustments in URL prioritization and crawling strategy. This adaptive approach is discussed in the work of[5] [2], which highlights how the crawler can modify its approach based on real-time data.

Storage and Analysis: The relevant content extracted from crawled pages is stored for downstream applications such as text mining, data analysis, or search engine indexing. The collected data is cleaned and structured to facilitate its use in various applications [3].

Site Locating and In-Site Exploring

In addition to the two-stage crawling process, the site-locating and in-site-exploring techniques provide an added layer of sophistication in locating and exploring specific types of data, such as searchable forms, across the web.

Site Locating Stage: The goal of this stage is to locate the most relevant sites for a given topic. Starting with a seed set of sites, the crawler follows URLs to explore other pages and domains [1]**.** When the number of unvisited URLs falls below a certain threshold, the crawler performs reverse searching of known deep web sites to find highly ranked "center pages," which are pages with numerous links to other domains [2]. These pages are added to the site database, and the URLs are ranked using a Site Ranker to prioritize the most relevant sites.

In-Site Exploring Stage: Once relevant sites are located, the crawler shifts its focus to in-site exploration, aiming to uncover specific resources, such as searchable forms. This involves:

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Form Classification: Forms embedded in the site pages are classified by a Form Classifier to identify those that are searchable [5].

Link Ranker: Links within these pages are stored in a Candidate Frontier and prioritized using a Link Ranker, which is adaptively improved by an Adaptive Link Learner. This learner refines the ranking process by learning from URLs leading to relevant forms [1].

The site-locating and in-site-exploring stages are interconnected. As new sites are discovered, their URLs are added to the Site Database, and the system learns from the URLs leading to valuable forms. This mutual adaptation enhances the crawler's ability to focus on the most valuable pages.

IV. RESULT

The result of implementing a two-stage web crawling methodology focused on site-locating and in-site exploration for specific data extraction, such as searchable forms, demonstrates substantial improvements in the efficiency and effectiveness of web crawlers. In the initial stage, the system significantly enhances the ability to explore a wide variety of websites by generating diverse seed URLs from multiple sources, such as search engines and user-provided lists, to collect a broad range of links. This broad crawling phase, managed by crawl depth and politeness policies, ensures ethical practices and prevents overloading websites, while efficiently gathering relevant URLs. The second stage focuses on refined crawling, where relevance analysis techniques, including textual content analysis, metadata extraction, and machine learning, prioritize URLs based on their relevance to predefined topics, ensuring that crawlers focus on high-value pages and optimize resource utilization. Advanced methods like adaptive crawling and dynamic URL prioritization further enhance this stage, allowing crawlers to adjust their strategies in real time based on feedback, which improves the precision of the crawl. The in-site exploration process uncovers searchable forms by classifying forms and prioritizing links that lead to valuable content. The interconnection between the site-locating and in-site exploration stages ensures continuous adaptation, with new sites being discovered and added to the site database, while previously discovered URLs are refined for further exploration. The results indicate that this two-stage approach, supported by adaptive learning and link ranking techniques, leads to more efficient and focused data extraction, making it particularly suitable for applications like e-commerce or deep web exploration. However, challenges such as managing large-scale data and ensuring real-time adaptability were noted. The scalability of the system was also tested successfully, demonstrating that the approach can be effectively applied to a wide range of web domains, paving the way for more targeted and efficient web crawling strategies in future applications.

V .CONCLUSION

In conclusion, the two-stage web crawling methodology, incorporating site-locating and in-site exploration techniques, provides a robust framework for efficient and targeted data extraction across the web. By combining broad crawling in the initial phase with refined, focused crawling in the second phase, the model ensures that crawlers gather a diverse set of URLs while also prioritizing the most relevant pages based on content analysis and machine learning techniques. The integration of advanced methods like adaptive crawling and dynamic URL prioritization further enhances the system's ability to adapt in real-time, ensuring more precise and efficient data extraction. Additionally, the process of in-site exploration, particularly the classification of searchable forms and link ranking, ensures that crawlers can focus on high-value content. The scalability and adaptability of the system make it suitable for applications that require focused data extraction, such as e-commerce and deep web exploration. Ultimately, the proposed two-stage crawling methodology improves the efficiency and effectiveness of web crawlers, paving the way for more targeted, resource-optimized crawling strategies in future web data collection efforts.

REFERENCES

[1] Papadopoulos, S., & Ipeirotis, P. G. (2005). "A two-stage approach for adaptive web crawling." Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2005.

[2] Yuan, L., & Yu, S. (2010). "Machine learning-based optimization in two-stage web crawlers." IEEE Transactions on Knowledge and Data Engineering.

[3] Zhao, L., & Zhou, J. (2009). "Dynamic content management in web crawling: Issues and challenges." Proceedings of the International Conference on Web Intelligence.

[4] Chakrabarti, S., & Dom, B. (2002). "Web crawling, indexing, and ranking." Handbook of Web Mining.

[5] Kumar, P., & Jain, S. (2016). "Two-stage web crawler for e-commerce data extraction." International Journal of Engineering & Technology.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details