INTERNATIONAL ACADEMIC CONFERENCE ON DEVELOPMENT IN SCIENCE, ENGINEERING, TECHNOLOGY AND MANAGEMENT (IACDSETM – 2026)

New Delhi, India on 8th August, 2026

CERTIFICATE NO : IACDSETM /2026/C0826813

A Study of Machine Learning-Based Clustering Techniques for Optimizing Web Crawling and Search Engine Indexing

Pasnur Deeplaxmi Purshottam

Research Scholar, Ph. D. in Computer Science and Engineering, Mansarovar Global University, Sehore, M.P., India.

ABSTRACT

The continuous growth of web content has created challenges in discovering, organizing, and retrieving relevant information from large numbers of web pages. Traditional web crawling methods may consume considerable time and computational resources when identifying useful pages and updating search engine indexes. Machine learning-based clustering offers an effective approach by grouping web pages according to similarities in their content, structure, keywords, and link relationships. Techniques such as K-means clustering, hierarchical clustering, DBSCAN, and spectral clustering can help organize web pages into meaningful categories. These clusters can guide web crawlers toward relevant pages, reduce unnecessary crawling, and improve the prioritization of frequently updated or information-rich websites. Furthermore, clustering can support search engine indexing by organizing documents into related groups, reducing redundancy, and facilitating faster information retrieval. The integration of natural language processing and machine learning can further improve the identification of semantic relationships between web pages. However, these approaches face challenges related to large-scale data processing, dynamic web content, algorithmic complexity, and changing user information needs. Therefore, selecting appropriate clustering algorithms and evaluating their performance are essential for achieving reliable results. This study examines innovative machine learning-based clustering techniques and their potential to enhance crawling efficiency, indexing accuracy, resource utilization, and the overall performance of modern search engines.