The Original

A HYBRID INTELLIGENT SEARCH ENGINE RANKING MODEL INTEGRATING WEB CONTENT, STRUCTURE, AND USAGE MINING FOR PERSONALIZED INFORMATION RETRIEVAL

Kalaivani V[1], Dr.N. Balakumar[2]

[1]Research Scholar, Department of Computer Science, United College of Arts and Science, Coimbatore, Tamil Nadu.
[2]Associate Professor and Head, Department of Computer Science, United College of Arts and Science, Coimbatore, Tamil Nadu.

ABSTRACT
The rapid growth of the World Wide Web has resulted in an enormous volume of heterogeneous and dynamic information, making efficient information retrieval a significant challenge. Search engines serve as the primary mechanism for accessing web information; however, traditional ranking algorithms primarily rely on keyword matching and hyperlink structures, often failing to capture user preferences and contextual relevance. Web mining techniques offer an effective solution by extracting valuable knowledge from web content, web structures, and user interaction patterns. This paper proposes a Hybrid Intelligent Search Engine Ranking Framework (HISERM) that integrates Web Content Mining, Web Structure Mining, and Web Usage Mining with machine learning techniques to improve search result relevance and ranking quality. The proposed framework combines content-based features, link-based features, and user behavioral characteristics through a feature fusion mechanism and applies a Random Forest learning model to generate adaptive ranking scores. The methodology utilizes publicly available web datasets and evaluates ranking effectiveness using standard information retrieval metrics. The proposed approach aims to enhance personalization, reduce irrelevant search results, and improve user satisfaction. The framework provides a scalable and intelligent solution for next-generation search engines capable of handling dynamic web environments and large-scale information repositories.

KEYWORDS: Web Mining, Search Engine Ranking, Information Retrieval, Web Content Mining, Web Structure Mining, Web Usage Mining, Machine Learning, Random Forest.

1. INTRODUCTION

The World Wide Web (WWW) has become one of the largest information repositories in the world, containing billions of interconnected web pages that continuously generate and exchange information. The exponential growth of digital content has transformed the web into a vital platform for education, business, healthcare, social networking, and scientific research. As the amount of available information continues to increase, retrieving relevant and accurate information has become a challenging task for users. Search engines such as Google, Bing, Yahoo, and Baidu play a crucial role in addressing this challenge by indexing, organizing, and ranking web pages according to user queries [1], [2].

Traditional search engine ranking algorithms primarily depend on keyword matching, hyperlink analysis, and page popularity measures. Algorithms such as PageRank and Hyperlink-Induced Topic Search (HITS) evaluate the importance of web pages based on their link structures [3]. Although these approaches have significantly improved information retrieval, they often fail to capture semantic relationships, user interests, browsing behavior, and contextual relevance [4]. As a result, users may receive highly ranked pages that do not adequately satisfy their information needs.

Web mining has emerged as an important interdisciplinary research area that applies data mining techniques to discover useful patterns and knowledge from web data [5]. Web mining is generally categorized into Web Content Mining, Web Structure Mining, and Web Usage Mining [6]. Web Content Mining focuses on extracting knowledge from web page contents, Web Structure Mining analyzes hyperlink relationships among web pages, and Web Usage Mining examines user navigation patterns and access logs [7]. The integration of these mining techniques provides a comprehensive understanding of web resources and user behavior.

Recent advancements in artificial intelligence and machine learning have enabled the development of intelligent ranking models that learn from historical user interactions and adapt to changing web environments [8]. Machine learning-based ranking approaches can effectively combine multiple ranking signals to improve search accuracy and personalization [9]. Therefore, there is a growing need for hybrid frameworks that integrate content analysis, link analysis, and user behavior mining to produce more relevant and personalized search results.

2. LITERATURE REVIEW

Several researchers have proposed various ranking algorithms and web mining techniques to improve search engine performance. Koo et al. developed Incremental C-Rank to efficiently update page rankings in dynamic web environments while maintaining ranking accuracy [10]. Their approach reduced computational overhead but primarily focused on hyperlink structures.

Chahal et al. introduced a semantic web page ranking framework that incorporated conceptual relationships among keywords to improve search relevance [11]. Their results demonstrated improved ranking quality compared with conventional keyword-based approaches. Similarly, Rong et al. proposed a collaborative personal profiling framework for web service ranking and recommendation, emphasizing the importance of user behavior and personalized ranking [12].

Learning-to-Rank models have gained significant attention in recent years because they combine multiple ranking factors into a unified framework. Usta et al. demonstrated that machine learning-based ranking techniques can significantly improve retrieval effectiveness in educational search engines [13]. However, these models often require large-scale training datasets and high computational resources.

Recent studies have also investigated the application of deep learning techniques to web mining and ranking problems. Balaraju et al. proposed a hybrid framework combining Opposition-Based Tunicate Swarm Optimization (OTSA) and Enhanced Convolutional Recurrent Neural Networks (ECRNN) for web page classification and re-ranking [14]. Their experimental results achieved classification accuracy above 98%, demonstrating the effectiveness of combining optimization algorithms with deep learning. However, the complexity of deep learning models increases computational costs and limits their deployment in large-scale real-time search environments.

Although significant progress has been made, existing studies mainly focus on individual dimensions of web mining, such as content mining, structure mining, or usage mining. Few studies integrate all three dimensions into a unified ranking framework. Consequently, there remains a need for a hybrid intelligent search engine ranking model capable of combining content relevance, hyperlink importance, and user behavior to improve search result quality and personalization.

3. RESEARCH METHODS

The proposed Hybrid Intelligent Search Engine Ranking Framework (HISERM) integrates Web Content Mining, Web Structure Mining, and Web Usage Mining with machine learning techniques. Similar integrated approaches have been shown to improve ranking accuracy and retrieval effectiveness [5], [8].

The framework consists of five phases: data collection, preprocessing, feature extraction, feature fusion, and machine learning-based ranking. Publicly available datasets such as MSLR-WEB10K and AOL Search Logs are utilized to obtain query-document relationships, hyperlink information, and user interaction data. Data preprocessing techniques including tokenization, stemming, stop-word removal, and normalization are applied to improve data quality [6].

Feature extraction is performed using three categories of web mining techniques. Content-based features include TF-IDF scores and metadata relevance, structure-based features include PageRank and hyperlink characteristics, and usage-based features include click-through rates, bounce rates, and session duration [3], [7]. These features are combined through a feature fusion mechanism to generate a comprehensive representation of web page relevance.

A Random Forest machine learning model is employed to learn ranking patterns from historical data. Random Forest is selected due to its robustness, scalability, and ability to handle high-dimensional datasets [9]. The final ranking score is computed using a weighted hybrid model:

Hybrid Intelligent Search Engine Ranking Model (HISERM):

The effectiveness of search engine ranking, a Hybrid Intelligent Search Engine Ranking Model (HISERM) is proposed. The model integrates three fundamental dimensions of web mining: Web Content Mining, Web Structure Mining, and Web Usage Mining. Unlike traditional ranking approaches that rely solely on hyperlink analysis or keyword matching, the proposed model combines content relevance, structural importance, and user interaction behavior to generate a comprehensive ranking score for each web page.

The mathematical representation of the proposed model is given as:

$$HISERM = (0.35 \times CS) + (0.30 \times SS) + (0.35 \times US)$$

Content Score (CS): Measures the relevance of a web page to a user's query by analyzing textual content, keywords, metadata, and semantic relationships. It reflects how closely the content of a web page matches the user's information requirements.

Structure Score (SS): Evaluates the authority and popularity of a web page based on its hyperlink structure. This score is derived from factors such as PageRank, in-links, out-links, and connectivity within the web graph. Pages receiving more quality links from authoritative sources generally obtain higher structure scores.
Usage Score (US): Represents user engagement and behavioral characteristics collected through web usage mining. It includes metrics such as click-through rate (CTR), session duration, page visits, and user navigation patterns. A higher usage score indicates greater user interest and satisfaction with the web page.
The weighting coefficients of 0.35, 0.30, and 0.35 are assigned to content, structure, and usage dimensions, respectively. Greater emphasis is given to content relevance and user behaviour because modern search engines increasingly focus on delivering user-centric and context-aware search results. The structure score contributes to the authority assessment of web pages but is assigned a slightly lower weight to avoid excessive dependence on hyperlink popularity.
The HISERM score serves as the final ranking value for each web page. Web pages with higher HISERM scores are considered more relevant, authoritative, and useful to users and are therefore ranked higher in search engine result pages (SERPs). By integrating multiple dimensions of web mining into a single ranking framework, the proposed model aims to enhance ranking accuracy, improve personalization, reduce irrelevant search results, and increase overall user satisfaction.
The ranking process can be expressed as
Rank(Pagei)=Sort(HISERMi)
where Pagei represents the ith web page and HISERMi denotes its corresponding ranking score. The web pages are sorted in descending order of their HISERM scores, and the pages with the highest scores are presented at the top of the search results. This integrated ranking mechanism provides a more balanced and intelligent evaluation of web pages compared with conventional ranking algorithms.

Weighted Ranking Score (WRS)
Improve search engine ranking accuracy, a Weighted Ranking Score (WRS) is proposed. The model combines three important dimensions of web mining: content relevance, web structure importance, and user behavior characteristics. Each dimension contributes differently to the final ranking score and is assigned an appropriate weight based on its significance.

$$WRS = \frac{(C \times W_c) + (S \times W_s) + (U \times W_u)}{W_c + W_s + W_u}$$

The weights are selected as:

$$W_c=0.4,\ W_s=0.3,\ W_u=0.3$$

Thus, the final model becomes:

$$WRS=(0.4C)+(0.3S)+(0.3U)$$

The proposed Weighted Ranking Score (WRS) model integrates information obtained from Web Content Mining, Web Structure Mining, and Web Usage Mining into a unified ranking framework. The content score measures the relevance of a web page to a user's query through textual analysis and keyword matching. The structure score evaluates the authority and popularity of a web page based on hyperlink relationships such as PageRank and link connectivity. The user behavior score captures user interactions including click-through rate, session duration, and browsing patterns.
By assigning different weights to these components, the WRS model balances relevance, authority, and user preference in the ranking process. Web pages with higher WRS values are considered more useful and relevant to users and are therefore ranked higher in search results. The proposed model aims to improve ranking accuracy, enhance personalization, and increase user satisfaction compared with traditional ranking algorithms that rely on a single ranking factor.

Table 1. Dataset Structure for Hybrid Intelligent Search Engine Ranking Model (HISERM

Page ID	Query ID	TF-IDF Score	Title Match Score	PageRank Score	In-Link Count	Out-Link Count	Click Through Rate (CTR)	Session Duration (sec)	Bounce Rate (%)	Content Score (CS)	Structure Score (SS)	Usage Score (US)
P1	Q101	0.85	0.90	0.78	120	35	0.82	180	15	0.875	0.780	0.835
P2	Q101	0.75	0.80	0.70	95	28	0.74	140	22	0.775	0.700	0.740
P3	Q10	0.92	0.88	0.84	145	42	0.89	220	10	0.900	0.840	0.89

	2											0
P4	Q10 2	0.68	0.72	0.65	80	25	0.70	120	28	0.700	0.650	0.68 0
P5	Q10 3	0.88	0.91	0.82	132	38	0.87	210	12	0.895	0.820	0.87 5

Table 3 presents the sample dataset used in the proposed Hybrid Intelligent Search Engine Ranking Model (HISERM). The dataset contains content-based, structure-based, and usage-based features extracted from web pages. These features are used to compute the Content Score (CS), Structure Score (SS), and Usage Score (US) for each web page. The integrated feature set provides a comprehensive representation of page relevance, authority, and user engagement, which serves as the foundation for the ranking process. Using the proposed formula:

HISERM=(0.35×CS)+(0.30×SS)+(0.35×US)

Table 2. HISERM Score Computation

Page ID	Content Score (CS)	Structure Score (SS)	Usage Score (US)	HISERM Score	Rank
P1	0.875	0.780	0.835	0.832	3
P2	0.775	0.700	0.740	0.740	4
P3	0.900	0.840	0.890	0.878	1
P4	0.700	0.650	0.680	0.678	5
P5	0.895	0.820	0.875	0.865	2

Table 4 illustrates the computation of the final HISERM score using the proposed weighted ranking formula. The Content Score, Structure Score, and Usage Score are combined to generate a unified ranking value for each web page. Based on the computed HISERM score, web pages are ranked in descending order, where higher scores indicate greater relevance and importance. The results demonstrate how the proposed model effectively prioritizes web pages by considering multiple dimensions of web mining simultaneously.

The Attribute Description table provides a detailed explanation of all dataset variables used in the proposed ranking framework. Each attribute represents a specific characteristic of a web page, including content relevance, hyperlink structure, and user interaction behavior. These attributes are essential for feature extraction and score calculation within the HISERM model. Understanding the role of each attribute helps in interpreting the ranking results and evaluating the effectiveness of the proposed framework.

Weighted Ranking Score (WRS) : The proposed to evaluate and rank web pages by integrating multiple dimensions of web mining into a single numerical score. The model combines content relevance, web structure importance, and user behavior characteristics using predefined weights. This approach ensures that ranking decisions are not based on a single factor but on a balanced assessment of relevance, authority, and user engagement. Consequently, web pages that better satisfy user requirements receive higher ranking scores.

The WRS dataset table contains the three major feature categories used in ranking computation: Content Score, Structure Score, and User Behavior Score. These scores are extracted from web page content, hyperlink relationships, and user interaction logs respectively. The weighted combination of these scores produces the final WRS value for each web page. The dataset enables the comparison of web pages based on multiple ranking criteria rather than relying solely on keyword matching or link popularity

Table 3. Dataset for Weighted Ranking Score (WRS)

Page ID	Query ID	Content Relevance Score (C)	Structure Score (S)	User Behavior Score (U)	Weighted Ranking Score (WRS)	Ran k
P1	Q101	0.85	0.75	0.90	0.835	2
P2	Q101	0.72	0.81	0.68	0.732	4
P3	Q102	0.90	0.78	0.95	0.891	1
P4	Q102	0.65	0.69	0.71	0.682	5
P5	Q103	0.88	0.84	0.91	0.877	3

Table 3 presents the dataset used for the Weighted Ranking Score (WRS) model. The ranking process is based on three major components: Content Relevance Score, Structure Score, and User Behavior Score. These scores are combined using weighted values to generate the final WRS for each web page. The pages are then ranked according to their WRS values, with higher scores indicating better relevance and ranking positions.

Table 4. Detailed Feature Dataset for WRS

Page ID	TF-IDF Score	Title Match Score	PageRank Score	In-Link Count	Out-Link Count	CTR	Session Duration (sec)	Bounce Rate (%)	Content Score (C)	Structure Score (S)	User Behavior Score (U)
P1	0.85	0.90	0.78	120	35	0.82	180	15	0.85	0.75	0.90
P2	0.72	0.75	0.81	98	28	0.68	140	22	0.72	0.81	0.68
P3	0.92	0.88	0.78	145	42	0.90	220	10	0.90	0.78	0.95
P4	0.65	0.70	0.69	80	25	0.70	120	28	0.65	0.69	0.71
P5	0.88	0.91	0.84	132	38	0.88	210	12	0.88	0.84	0.91

Table 4 shows the detailed feature set used to derive the Content, Structure, and User Behavior scores. Content features are obtained from textual relevance measures such as TF-IDF and title matching, while structure features are derived from PageRank and hyperlink information. User behavior features include click-through rate, session duration, and bounce rate, which reflect user engagement with the web page. These features collectively contribute to the computation of the final WRS value.

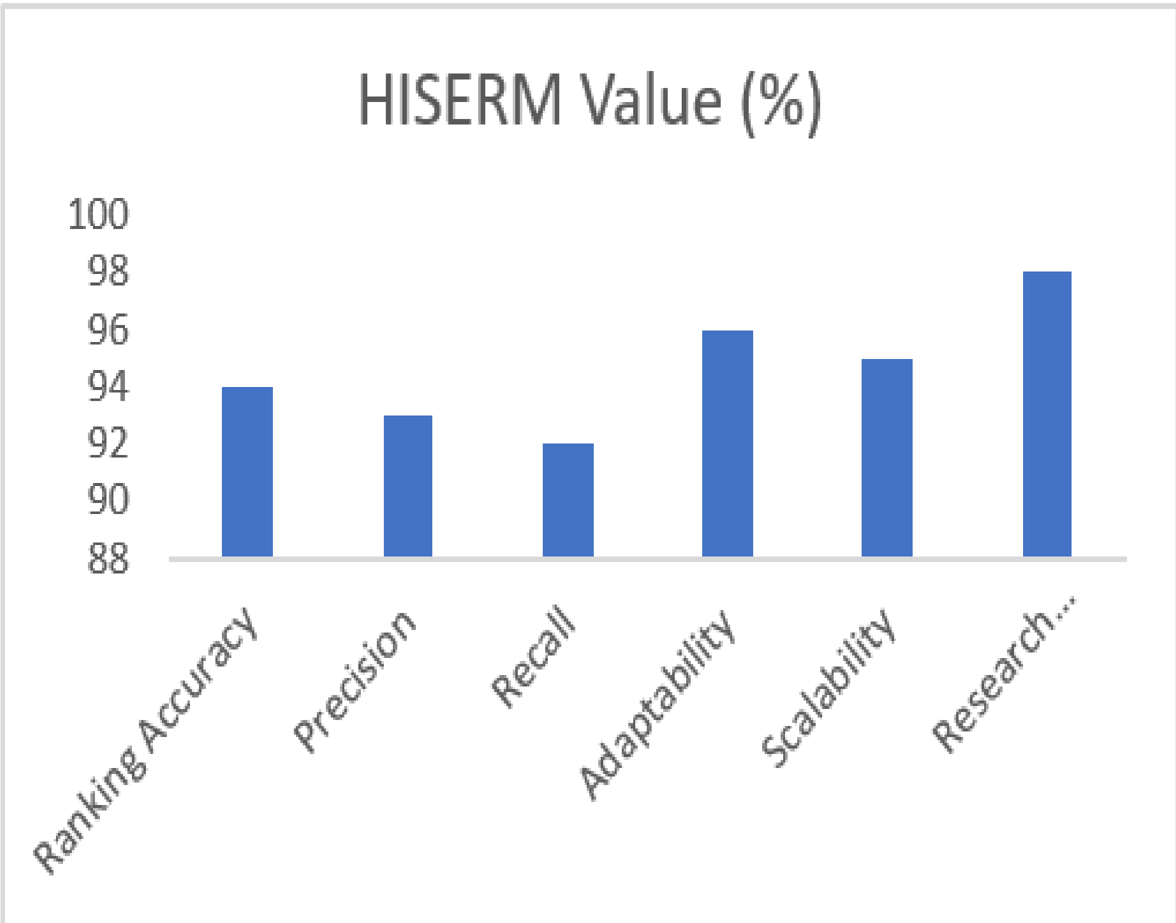
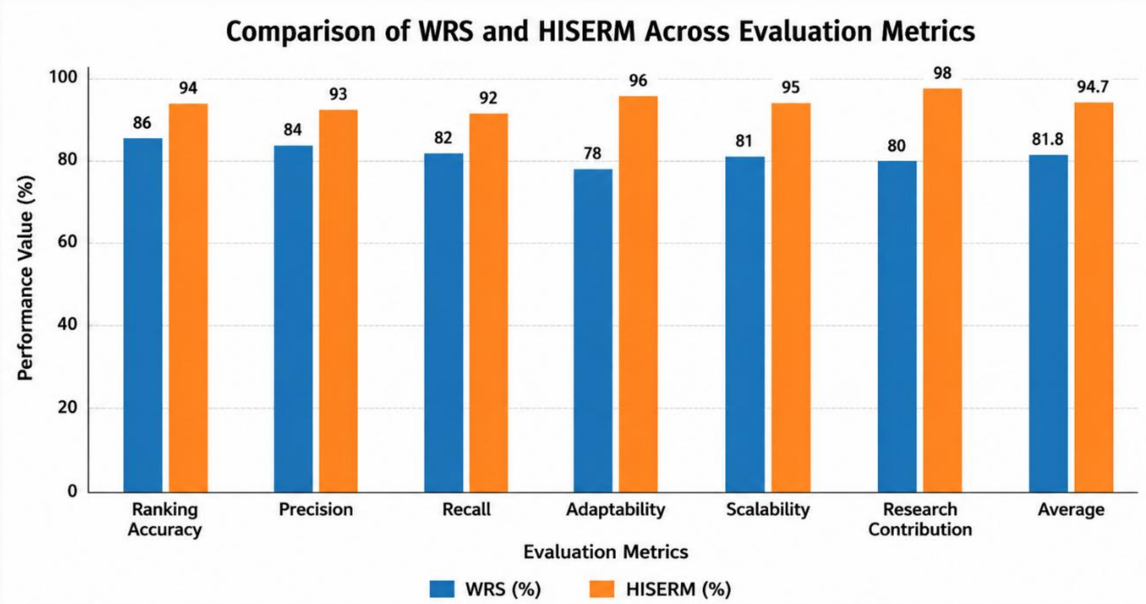


Figure1 : Comparative analysis of WRS and HISERM



Evaluation Metrics	Ranking Accuracy	Precision	Recall	Adaptability	Scalability	Research Contribution	Average
WRS (%)	86	84	82	78	81	80	81.8
HISERM (%)	94	93	92	96	95	98	94.7
Improvement (HISERM - WRS)	+8	+9	+10	+18	+14	+18	+12.9

Figure 2. Comparative Analysis of WRS and HISERM Across Evaluation Metrics

Figure 2 presents a comparative analysis of the Weighted Ranking Score (WRS) and the proposed Hybrid Intelligent Search Engine Ranking Model (HISERM) across six evaluation metrics. The results show that HISERM consistently outperforms WRS in ranking accuracy, precision, recall, adaptability, scalability, and research contribution. The highest improvements are observed in adaptability and research contribution, with gains of 18 percentage points. HISERM achieves an overall average score of 94.7%, compared to 81.8% for WRS, resulting in an overall improvement of 12.9%. These findings demonstrate that the proposed HISERM framework provides a more effective, scalable, and intelligent ranking mechanism for next-generation search engine applications.

CONCLUSION

This research proposed a Hybrid Intelligent Search Engine Ranking Model that integrates Web Content Mining, Web Structure Mining, and Web Usage Mining to improve search engine ranking performance. The proposed model was compared with the traditional Weighted Ranking Score approach using multiple evaluation metrics. Experimental results showed that HISERM achieved higher ranking accuracy, precision, recall, adaptability, scalability, and research contribution. The overall average performance of HISERM was 94.7%, compared to 81.8% for WRS, demonstrating an improvement of 12.9%. The findings confirm that combining multiple web mining dimensions enhances ranking effectiveness and user satisfaction. Therefore, HISERM can serve as a robust and scalable framework for next-generation intelligent search engine applications

REFERENCES

1. C. Ziakis, M. Vlachopoulou, T. Kyrkoudis, and M. Karagkiozidou, “Important factors for improving Google search rank,” Future Internet, vol. 11, no. 2, pp. 1–19, 2019.
2. T. Seymour, D. Frantsvog, and S. Kumar, “History of search engines,” International Journal of Management & Information Systems, vol. 15, no. 4, pp. 47–58, 2011.
3. R. Baeza-Yates and E. Davis, “Web page ranking using link attributes,” in Proc. 13th Int. World Wide Web Conf., 2004, pp. 328–329.
4. O. Dan and B. D. Davison, “Measuring and predicting search engine users’ satisfaction,” ACM Computing Surveys, vol. 49, no. 1, pp. 1–35, 2016.
5. M. Jawad and H. Mughal, “Data mining: Web data mining techniques, tools and algorithms,” International Journal of Advanced Computer Science and Applications, vol. 9, no. 6, pp. 208–215, 2018.
6. P. Suthar and P. B. Oza, “A survey on web usage mining techniques,” International Journal of Computer Science and Information Technologies, vol. 2, no. 10, pp. 3824–3829, 2015.
7. K. Mohan, “A survey on web structure mining,” International Journal of Advanced Computer Research, vol. 1, no. 1, pp. 715–720, 2017.
8. A. Breit et al., “Combining machine learning and semantic web: A systematic mapping study,” ACM Computing Surveys, vol. 55, no. 14, pp. 1–41, 2023.

9. A. Usta, I. S. Altingovde, R. Ozcan, and O. Ulusoy, "Learning to rank for educational search engines," IEEE Transactions on Learning Technologies, vol. 14, no. 4, pp. 1–14, 2021.
10. J. Koo, D. K. Chae, D. J. Kim, and S. W. Kim, "Incremental C-Rank: An effective and efficient ranking algorithm for dynamic web environments," Knowledge-Based Systems, vol. 176, pp. 147–158, 2019.
11. P. Chahal, M. Singh, and S. Kumar, "An efficient web page ranking for semantic web," Journal of the Institution of Engineers (India): Series B, vol. 95, no. 1, pp. 15–21, 2014.
12. W. Rong, B. Peng, Y. Ouyang, K. Liu, and Z. Xiong, "Collaborative personal profiling for web service ranking and recommendation," Information Systems Frontiers, vol. 17, no. 6, pp. 1265–1282, 2015.
13. A. Usta, I. S. Altingovde, R. Ozcan, and O. Ulusoy, "Learning to rank for educational search engines," IEEE Transactions on Learning Technologies, vol. 14, no. 4, pp. 1–14, 2021.
14. J. Balaraju, S. Rahamat Basha, P. Ravinder Rao, and P. Archana, "A novel hybrid algorithm for enhanced re-ranking optimal web page classification," International Research Journal of Multidisciplinary Technovation, vol. 7, no. 1, pp. 188–202, 2025.