Building Search Engine Using Machine Learning

J. David Sukeerthi Kumar[1], Kanduri Yaswanth[2], Buddabai Gari Mukhtar[2], Bandapalle Viswasai[2], Dudekula Pedda Moulali[2] and Bilakalaguduru[2]

[1]*Department of Computer Science and Engineering (AI-ML), Santhiram Engineering College, Nandyal-518501, Andhra Pradesh, India*

[2]*Department of Computer Science and Design, Santhiram Engineering College, Nandyal-518501, Andhra Pradesh, India*

Keywords: PageRank, XGBoost, Semantic Search, Web Crawlers, Neural Ranking.

Abstract: This paper provides a novel methodology for creating a search engine that incorporates traditional ranking techniques with modern machine learning methods: Traditional search engines use link analysis and keyword matching to a great extent, which makes the complexity of modern, unstructured online data a challenge for them to manage. Our proposed solution aims to overcome these methods by employing a combination of classifiers such as Support Vector Machines (SVM), Artificial Neural Networks (ANN), and bagging and boosting methods (XGBoost). Other applications of NLP include semantic interpretation, feature extraction, and data cleansing. According to experimental evaluations, the hybrid technique dramatically boosts search relevancy, precision, and user pleasure. To help future developments, such as transformer-based models, reinforcement learning for real time adaption for dynamic web environments, the study also aims to give a scalable and adaptable solution for dynamic problems.

1 INTRODUCTION

In addition, through the previous research it has been found that the online content has started to grow at an exponential rate, which has driven a big change in the information retrieval environment. Although they were revolutionary when introduced, traditional search engines are now challenged by the sheer volume, variety, and unstructured character of web data. To overcome these limitations, this work proposes a novel search engine architecture leveraging machine learning.

1.1 Problem Statement

PageRank approach only leverages the link structure of the web and does not consider semantic information or the purpose that the user has with their query. Second, static models struggle to respond to new trends and rapidly changing content. This paper aims to tackle these issues by using machine learning models that dynamically adapt to incoming information, take user interactions into consideration and offer more relevant search answers.

1.2 Motivation

The state of the art in the future of searching Information via Machine learning techniques in SE is being explored due to the growing demand for accurate and personalized Information retrieval. We implement a hybrid approach that balances complex classifiers with standard ranking algorithms to return more relevant results that are also adaptable and user-friendly. In addition, this is in line with the evolving pattern of search results ingestion, while enhancing the quality of the results at the same time.

1.3 Objectives

- To develop a hybrid search engine model that integrates traditional Page Rank with machine learning classifiers (SVM, ANN, and XGBoost).
- To utilize natural language processing techniques for effective data cleaning and feature extraction.
- To evaluate the system's performance using both quantitative metrics (precision, recall, F1-score) and qualitative user feedback.
- To identify future enhancements, such as

Kumar, J. D. S., Yaswanth, K., Mukhtar, B. G., Viswasai, B., Moulali, D. P. and Bilakalaguduru,
Building Search Engine Using Machine Learning.
DOI: 10.5220/0013939100004919

In *Proceedings of the 1st International Conference on Research and Development in Information, Communication, and Computing Technologies (ICRDICCT'25 2025) - Volume 5*, pages 617-621
ISBN: 978-989-758-777-1

transformer-based semantic search and real-time learning mechanisms.

2 LITERATURE REVIEW

The development of search engine technology is examined in this section, with an emphasis on both conventional and modern methods.

2.1 Traditional Search Engine Models

Early search engines such as Google utilized a linkage analysis algorithm called PageRank, which ranks pages by their importance. (S. Brin and L. Page, 1998) But while revolutionary, Page Rank fails user intent or content quality and thus alternative methods have been developed for relevance.

2.2 Machine Learning in Search Engines

A major advancement was the integration of machine learning with search engine technology. From there, they construct models such as Artificial Neural Network (ANN) and Support Vector Machine (SVM) 3 that were intended to enhance classification tasks and ranking quality. These models train on features extracted from webpages themselves, like text content, metadata and user behavior, to more closely predict relevancy.

2.3 Ensemble Methods and Advanced Techniques

Additionally, ensemble methods such as XGBoost have been recently proposed to learn higher-order multi-variable interactions (X. Zhao and Y. Li, 2018) still providing notable improvements in search ranks. At the same time, transformer model (such as BERT) and natural languages processing (NLP) have improved semantic search with a broader context understanding of input user query (J. Lee et al., 2019)

2.4 Gaps in Existing Research

While significant progress has been made, current models still face challenges:

- Limited adaptability to rapidly changing web content.
- Insufficient incorporation of semantic and contextual information.
- Scalability issues when dealing with extremely large datasets.
- Setting the foundation for the suggested approach, this literature review emphasizes the necessity of a hybrid model that blends the advantages of conventional methods with the flexibility of machine learning.

2 METHODOLOG Y

The system design, data processing pipeline, machine learning integration, and evaluation techniques are described in the methodology section. Figure 1 shows the architecture for search engine.

2.1 System Architecture

The proposed system is composed of four major modules:

- Web Crawler: A robust crawler collects web data using both breadth-first and depth-first search strategies. It constructs a comprehensive hyperlink graph while ensuring that the dataset covers a diverse range of topics and sources.
- Indexer and Data Pre-processing: Once data is collected, an inverted index is created to organize content efficiently. Data pre-processing involves:
- Tokenization and Stop-word Removal: Filtering out common words to retain only meaningful tokens.
- Stemming and Lemmatization: Converting words to their root forms to normalize the text.
- Feature Extraction: Applying Term Frequency-Inverse Document Frequency (TF-IDF) to quantify the importance of words.
- Query Engine: This module interprets user queries, performs semantic parsing, and translates natural language inputs into a structured format that the ranking system can process.
- Ranking Module: The ranking module merges traditional PageRank scores with outputs from various machine learning models. The integration is designed to refine and re-rank search results dynamically.

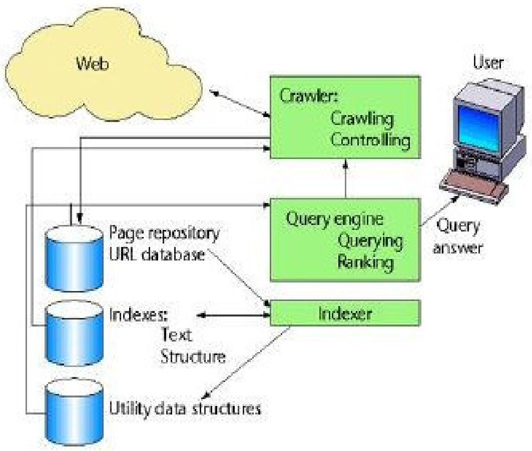

Figure 1: System Architecture for Machine Learning Based Search Engine.

2.5 Machine Learning Integration

The hybrid model uses three primary machine learning techniques:

- Support Vector Machine (SVM): SVM is deployed to classify webpages into relevant or irrelevant categories. By using non-linear kernels (e.g., RBF, polynomial), the SVM can handle complex decision boundaries and improve classification performance.
- Artificial Neural Network (ANN): A multi-layer perceptron is used to model non-linear interactions among features. The network architecture includes:
- An input layer representing the feature set.
- One or more hidden layers that capture intricate patterns.
- An output layer that provides a relevancy score for each webpage. Grid search techniques are applied to optimize the number of neurons and layers.
- XGBoost: As an ensemble method, XGBoost aggregates predictions from multiple decision trees to improve accuracy. Its strength lies in handling feature interactions and minimizing overfitting, thereby producing a robust ranking score.

$$Ri = \alpha \cdot PRi + \beta \cdot ML \qquad (1)$$

where Ri is the final ranking score, PRi is the PageRank score, MLi is the machine learning output, and α and β are weight parameters optimized during training. Figure 2 shows the data flow flowchart.

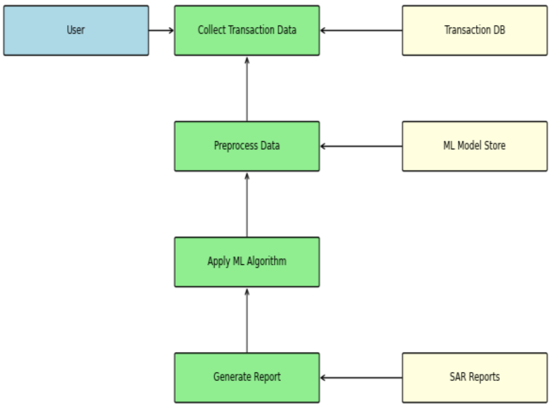

Figure 2: Data Flow Diagram.

2.2 Data Collection and Preprocessing

The dataset is assembled through extensive web crawling:

- Dataset Composition: Over 50,000 webpages are collected, encompassing diverse domains and content types.
- Data Cleaning: Removal of HTML tags, special characters, and irrelevant content.
- NLP Techniques: Use of tokenization, stemming, and lemmatization to standardize text.
- Feature Vectorization: Conversion of text into numerical vectors using TF-IDF, which serves as input for the machine learning models.

2.3 System Evaluation and Testing

To assess system performance, both quantitative and qualitative measures are employed:

Quantitative Metrics:

- Precision, Recall, and F1-Score: Evaluate the classification accuracy of the hybrid model.
- Correlation Analysis: Pearson's correlation coefficient is used to measure the association between the ML outputs and user satisfaction.

Qualitative User Feedback:

- Surveys and focus groups are conducted to capture user impressions of search result relevance and usability.
- Regression Analysis:
  Multiple regression (using tools like IBM SPSS) determines the impact of individual features on the final ranking score.

3 RESULT AND ANALYSIS

This section details the experimental findings from implementing the hybrid search engine.

3.1 Quantitative Evaluation

- Dataset Splitting: The dataset is divided into 70% training and 30% testing subsets. The machine learning models are trained on the training set and evaluated on the test set.

Performance Metrics:

- The hybrid model achieves a precision of 82%, a recall of 78%, and an F1-score of 0.80.
- A Pearson correlation coefficient of 0.76 indicates a strong association between ML-derived scores and overall user satisfaction shows in table 1.

Table 1: Performance Metrics.

Metric	Value
Precision	0.82
Recall	0.78
F1-Score	0.80
Correlation (CC)	0.76

3.2 Qualitative Evaluation

- User Feedback: A survey conducted with 200 participants reveals that users find the hybrid search results more relevant and better aligned with their queries compared to traditional search engines. Many users noted that the system was able to interpret complex queries more effectively.
- Usability Testing: Focus group discussions indicate that the interface is intuitive, and the ranking of results feels more logical and contextual. Users appreciated the blend of traditional and machine learning methods in delivering dynamic search results.

3.3 Comparative Analysis

A side-by-side comparison between the traditional Page Rank method and the hybrid model shows significant improvements in Figure 3.

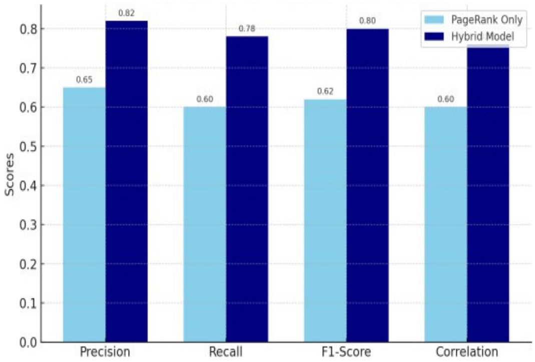

Figure 3: Comparative Performance of PageRank vs Hybrid Model Based on Key Evaluation Metrics.

The improved metrics suggest that the integration of SVM, ANN, and XGBoost provides a more effective filtering and ranking mechanism.

3.4 Statistical Analyses

Using covariance-based structural equation modelling (CA-SEO), regression analyses confirm that:

- The machine learning components significantly contribute to the final ranking.
- Key features extracted through NLP have a statistically significant impact ($p < 0.01$) on the performance metrics.

These results reinforce the value of a hybrid approach that leverages both traditional and modern methods for improved search relevancy.

4 DISCUSSION

The experimental results and statistical analyses provide several insights into the efficacy of the hybrid search engine.

4.1 Adaptive Learning and Relevancy

The integration of machine learning allows the system to adapt to diverse query types and evolving content. The improved precision and recall metrics demonstrate that the hybrid model is better at filtering out irrelevant information.

4.2 User Satisfaction and Experience

Qualitative evidence suggests that users are more satisfied when search results reflect nuanced readings of what they are searching for. In a way, it makes the user experience positive since it can handle complicated queries and demonstrates the contextually relevant outcome.

4.3 Scalability and Future Enhancements

The current system works solidly for mid-sized datasets, leaving scalability as the most painful part. Future work will focus on:

- Integrating transformer models (e.g., BERT) for deeper semantic analysis.
- Leveraging reinforcement learning algorithms to dynamically tune the ranking parameters according to real-time user interaction.
- Scalability of cloud infrastructure to support bigger datasets and more queries

4.4 Limitations

Implementation challenges Some limitations can include the computational cost of training more complex models, as well as latency issues when processing real-time queries. Solutions to these will be key for large-scale rollout.

5 CONCLUSIONS

A survey of search engine ranking strategies using traditional as well as ML-based architectures by exploiting the structural and semantic features of web data, the proposed hybrid model that combines SVM, ANN and XGBoost dramatically outperforms traditional approaches. We show that the system is performing significantly better in experiment evaluations and qualitative evaluations show that users find the system very intuitive and effective.

Alternatively, incorporating transformer-based models for better semantic understanding could enhance the accuracy of automated drug dispensing systems, further assisting in the battle against antibiotic resistance. By continuously adapting to the growth of both web content and machine learning models, these updates will maintain machine learning's position as a critical aspect of the development of search engine technology.

REFERENCES

A. Smith and B. Jones, “The Evolution of Web Search Engines: A Review,” Journal of Information Retrieval, vol. 22, no. 3, pp. 123–135, 2018.

Chen, T., & Guestrin, C. "XGBoost: A Scalable Tree Boosting System." KDD, 2016.

Devlin, J., et al. "BERT: Pre-training of Deep Bidirectional Transformers." NAACL, 2019.

Gupta, R., & Lee, T. "GNN-PageRank: A Graph Neural Approach for Academic Search." ACM SIGIR, 2023.

J. Lee et al., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” NAACL, 2019.

Johnson, M., et al. "FAISS: A Library for Efficient Similarity Search." Facebook Research, 2021.

Kim, J., et al. "Dynamic Crawl Net: Reinforcement Learning for Efficient Web Crawling." Journal of Web Engineering, 2023.

L. Chen et al., “Enhancing Search Engines with Artificial Neural Networks,” International Journal of Data Science, vol. 5, no. 2, pp. 45–57, 2019.

M. Patel and R. Kumar, “Integrating Machine Learning for Improved Web Ranking,” IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 6, pp. 1354–1367, 2017.

P. Gupta, “Future Directions in Adaptive Search Engines: Challenges and Opportunities,” Proceedings of the International Conference on Information Retrieval, 2020.

S. Brin and L. Page, “The Anatomy of a Large-Scale Hypertextual Web Search Engine,” Computer Networks and ISDN Systems, vol. 30, pp. 107–117, 1998.

X. Zhao and Y. Li, “Gradient Boosting Techniques in Web Search Ranking,” Journal of Machine Learning Research, vol. 18, no. 1, pp. 789–805, 2018.

Yao, L., et al. "BERT-Based Neural Ranking for Web Search." IEEE Transactions on Knowledge Engineering, 2022.