Boosting Webpage Retrieval with Ensemble Learning and Advanced Semantic Models: A Novel Re-Ranking Framework

Atul O. Thakare*, *Member, IAENG,* Narasimha Reddy Soora, Lambodar Jena, Arvind R. Singh

Abstract—**The exponential growth of web content has rendered traditional webpage retrieval techniques inadequate for addressing the evolving demands of modern users. This paper presents a novel framework that leverages advanced machine learning approaches to enhance retrieval effectiveness significantly. The proposed system integrates ensemble learning techniques—comprising Random Forests, Gradient Boosting, and Voting Classifiers—with transformer-based pre-trained language models such as BERT to capture deeper semantic representations. By jointly modeling syntactic and semantic similarity measures, the framework delivers highly relevant and context-aware search results. Empirical evaluations demonstrate that the proposed approach consistently outperforms conventional ranking methods, offering improved document relevance and higher user satisfaction across diverse query scenarios.**

Keywords: Webpage Retrieval, Page Rank algorithm, Page Re-ranking, Ensemble Learning, BERT, GPT, query expansion, feature extraction

I. Introduction

The rapid growth of web content has made the internet an essential global resource. The vast and ever-growing size of the content makes it challenging for search engines to provide accurate links to relevant web pages that align with the user's query. Google PageRank [1] and HITS [2] were revolutionary ranking methods when first introduced. These algorithms evaluate pages based on their link connections and relationship strengths. Early search engines ignored query semantics and user intent. Although Weighted PageRank [3] and EigenRumor [4] included blog influence and page popularity, they remained limited for content-based searches. To address these limitations, methods that take into account both user behavior and web page content have been proposed, such as content-based extensions [5] and time-aware ranking strategies [6]. But they still have a major flaw: they can't fully grasp semantic complexities. Automated [7], interactive [8], and hybrid query expansion techniques were developed to improve search relevance. These approaches effectively combine several methods to address the challenges of constructing improved queries. There are various strategies, and each one helps improve the efficiency and effectiveness of information retrieval systems. These techniques utilize linguistic tools such as WordNet, Word2Vec [9] and GloVe [10]. However, traditional methods for expanding user queries to achieve more accurate information retrieval have scalability and computational complexity issues, particularly with large datasets. The area of natural language processing (NLP) has been significantly reshaped by recent advancements in large language models (LLMs), such as BERT [11] and GPT [12]. Transformer models effectively extract syntactic and semantic text information. Self-attention in transformers [13] strengthens word context and semantics, improving search results. This capability supports various NLP tasks such as translation, summarization, and idea extraction. The use of these methods in information retrieval has created new opportunities for improving query interpretation and document ranking. We propose a new method combining ensemble learning and advanced semantic similarity to recommend relevant webpages. Ensemble methods, such as Random Forests [14], Gradient Boosting [15], and Voting Classifiers, are well-known for their robustness and ability to discover complex patterns in data, thereby improving predictive performance. However, when combined with semantic similarity information, the pre-trained transformer model significantly improves the quality of re-ranked search results. In addition to increasing the accuracy and quality of re-ranked search results, the proposed method is adaptable and scalable, providing relevant search results that correspond to various query intents. It is beneficial for matching user expectations and the utilization and throughput of the re-ranked search results. Thus, combining classical ML with LLMs improves information retrieval accuracy. Therefore, our research proposes a novel hybrid technology that aims to improve information retrieval results by combining the best features of machine learning techniques with large language models.

Manuscript received January 1, 2025; revised July 12, 2025

Atul O. Thakare is an Associate Professor in the School of Computing, MIT Art Design & Technology University, Pune, Maharashtra, India - 412201 (corresponding author to provide phone: +91 8767829219, e-mail: aothakare@gmail.com).

Narasimha Reddy Soora is a Professor and Head of the Department of CSE (AI & ML), in Kakatiya Institute of Technology & Science, Warangal, Telangana, India - 506015 (e-mail: snreddy75@gmail.com).

Lambodar Jena is a Professor in the Department of CSE, Siksha 'O' Anusandhan (Deemed to be) University, Bhubaneswar, Odisha, India - 751030 (e-mail: lambodarjena@soa.ac.in).

Arvind R. Singh is an Honorary Research Fellow at Applied Science Research Center in Applied Science Private University, Amman, Jordan - 11931 (e-mail: arvindsinghwce@gmail.com).

II. Literature Survey

Nogueira and Cho (2019) [16] cited the use of transformer-based models for re-ranking, and they demonstrated that BERT's contextual embeddings are more effective at capturing semantic relationships between queries and passages than traditional methods. Gao et al. (2021) [17] proposed SimCSE, a simple but effective framework for contrastive learning of sentence embeddings. SimCSE utilizes dropout, and contrastive learning can be effectively applied to sentence embeddings without requiring complex architectures or large amounts of labeled data. Zamani et al. (2022) [18] proposed an ensemble ranking framework to enhance existing web ranking systems. The work also underscores the importance of ensemble methods in information retrieval and highlights that combining different models (traditional and neural) yields more optimal and accurate ranking systems. Li et al. (2023) [19] provide a critical examination of the use of external knowledge in retrieval models. The authors identify multiple issues with current knowledge-augmented methods while proposing adaptive methods to incorporate learning into retrieval systems that focus on retrieval model adaptability and context awareness. Wu et al. (2023) [20] developed a ranking model that utilizes reinforcement learning to acquire knowledge from user interactions, dynamically modifying search relevance over time. The work demonstrates that reinforcement learning can be effective for search ranking by learning from user behavior in real-time. In addition, the authors note that one of the significant challenges is designing reward functions that accurately reflect user satisfaction and long-term engagement. Chen et al. (2024) [21] investigate the field of multimodal retrieval, which deals with retrieving information from diverse modalities, such as text, images, video, and audio. The work highlights the growing importance of multimodal retrieval in contemporary search systems, as users increasingly demand to search for and find information across multiple modalities. Furthermore, it highlights the need for rapid and scalable methods to handle multimodal data.

III. Problem Statement

With the increase in the size and complexity of web content, conventional ranking algorithms, such as PageRank and HITS, are inadequate because they are based on link-centric metrics that do not consider content relevance and user intent. Attempts, such as Weighted PageRank and EigenRumor, to incorporate the page popularity factor did not accurately represent semantic relationships between user queries and documents. The current query expansion methods, which refine search terms effectively, encounter difficulties with ambiguous queries and demonstrate computational inefficiency when processing large datasets. The semantic understanding of BERT and GPT models remains limited in ranking systems despite their advanced capabilities.

The main challenges in webpage retrieval include:

- **Semantic Gap:** Standard ranking methods produce results that are inappropriately relevant because they are unaware of the semantic details of the query and the document.
- **Scalability Issues:** Query expansion and ranking techniques are also computationally inefficient; therefore, they are not very practical for large-scale datasets.
- **Static Relevance Models:** The current methods employ static ranking models. Static ranking models are insensitive to dynamic factors, such as changing user preferences or differences in query intent.
- **Underutilization of Ensemble Learning:** Ensemble learning, a technique that improves model accuracy by combining multiple models, is not fully integrated into ranking methodologies. Using diverse similarity metrics and feature combinations might enhance search result relevance.

In this paper, a hybrid approach is proposed, which combines the strengths of ensemble methods and transformer models into a single model. The system also employs multiple similarity measures for performance improvement, including Euclidean Distance and Jensen-Shannon Divergence. Euclidean Distance quantifies the distance between two points in a given space, measuring numerical proximity for vector representation. At the same time, Jensen-Shannon Divergence computes probabilistic similarity between two distributions based on their overlap. Integration with geometric and probabilistic techniques helps optimize precision and relevance, yielding enhanced information retrieval results. Euclidean distance reveals the spatial relations of embeddings in space through vectors. The Jensen-Shannon divergence then provides a proper, in-depth look at semantic overlaps, which is especially well-suited for probabilistic embeddings.

IV. Proposed Solution

The proposed framework comprises three tightly integrated components: the system architecture, an ensemble learning-based ranking model, and a semantic similarity integration module. The solution aims to enhance scalability, adaptability, and retrieval relevance by combining advanced query expansion strategies with ensemble-based learning and state-of-the-art semantic similarity metrics.

A. System Architecture

The system architecture functions to handle user queries, extract webpage metadata and content features, and produce results that match user intent through ranking. The architecture consists of sequential stages, which start with the following core steps:

- **Query Preprocessing and Expansion:** To perform query expansion, the query is first preprocessed to remove stop words, normalize text, and tokenize terms.

Keyword suggestions are achieved using query expansion with the help of trained transformer models, such as BERT or GPT, to expand the query with contextually relevant keywords. Let the user query be referred to as Q. Then, the expanded query is represented as:

$$Q' = Q \cup \{k_1, k_2, \ldots, k_n\}$$

Where k_i represents additional keywords suggested by the transformer model.

- **Feature Extraction from Webpage Metadata and Content:** For each retrieved web, there are metadata (title, snippet, and URL) and content, which can be processed to extract features. Let metadata of webpage W be $M = \{m_1, m_2, m_3\}$ (title, snippet, URL), and content be C. Numerical representations are generated for M and C using word embeddings such as GloVe or embeddings from transformers:

$$\text{Embedding}(W) = \text{Encoder}(M, C)$$

Where Encoder is a neural network based transformer model.

- **Ensemble Model Training for Ranking and Reranking:** An ensemble model is a combination of base learners (e.g., Random Forests, Gradient Boosting) that are used to assign a ranking score to every webpage. The query-document pairs are first processed to extract features and then passed to the ensemble framework to predict the relevance score $R(W, Q)$ of each webpage W. These scores are then used to re-rank the web pages.

The overall system flow is depicted in Figure 1, which illustrates how query expansion, feature extraction, and the ensemble learning framework interact to improve the retrieval of webpages.

B. Ensemble Learning Framework

The system evaluates webpage relevance to user queries by using an ensemble learning framework that integrates multiple complementary models. The ensemble method enables the system to identify different data patterns and minimize the risks associated with using a single learning model. The framework combines the following models:

- **Random Forests:** The Random Forest algorithm builds multiple decision trees that use randomly selected webpage metadata features, along with user query features, for training. The trees in the ensemble vote to determine the relevance of the webpage. The Random Forest model calculates its final relevance score through an average prediction process across all trees. The relevance score is calculated as follows:

$$R_{\text{RF}}(W, Q) = \frac{1}{T} \sum_{t=1}^{T} R_t(W, Q)$$

where R_t is the predicted relevance score from the t^{th} tree, and T is the total number of trees.

- **Gradient Boosting:** The training process of Gradient Boosting differs from Random Forests because it sequentially constructs trees. The ensemble of previous learners receives correction attempts from each new tree that is added to it. The relevance score calculation for webpage W concerning query Q happens through an iterative process:

$$R_{\text{GB}}^{(i)}(W, Q) = R_{\text{GB}}^{(i-1)}(W, Q) + \eta \cdot h_i(W, Q)$$

where η is the learning rate, and h_i is the i_{th} weak learner.

- **Voting Classifier:** A Voting Classifier combines the strengths of multiple models by using weighted voting to aggregate predictions from Random Forests, Gradient Boosting, and other models. The final decision results from the weighted contributions of each model to the process:

$$R_{\text{Ensemble}}(W, Q) = \sum_{j=1}^{M} w_j \cdot R_j(W, Q)$$

where M is the number of models, w_j is the weight for model j and R_j is the relevance score from model j.

In this case, the final ensemble prediction is $R_{\text{Final}}(W, Q)$, using which the retrieved pages can be re-ranked, prioritizing those most relevant to the query.

C. Semantic Similarity Integration

The framework incorporates semantic similarity metrics for comparing the alignment between the query and webpage content in both geometric and probabilistic dimensions. Two critical similarity measures are used.

- **Euclidean Distance:** This Euclidean Distance is the geometric distance measurement between the query embedding $\vec{Q}$ and a webpage embedding $\vec{W}$:

$$S_{\text{Euclidean}}(Q, W) = \|\vec{Q} - \vec{W}\|_2$$

Implying where $\|\vec{Q} - \vec{W}\|_2$ denotes the L2 norm of difference.

- **Jensen-Shannon Divergence:** The Jensen-Shannon Divergence specifically measures the probability and similarity values between the query distributions Q and the webpage W:

$$S_{\text{JSD}}(Q, W) = 1 - \frac{1}{2}\left(D_{\text{KL}}(P\|M) + D_{\text{KL}}(Q\|M)\right)$$

where P and Q refer to the probability distributions of Q and W; here, $M = \frac{1}{2}(P + Q)$ is the averaged distribution, and D_{KL} is the Kullback-Leibler divergence.

- **Combined Similarity:** We will combine both geometric and probabilistic aspects to derive the final similarity score.

$$\begin{aligned} S_{\text{Combined}}(Q, W) = \beta \cdot (1 - S_{\text{Euclidean}}(Q, W)) \\ + (1 - \beta) \cdot S_{\text{JSD}}(Q, W) \end{aligned}$$

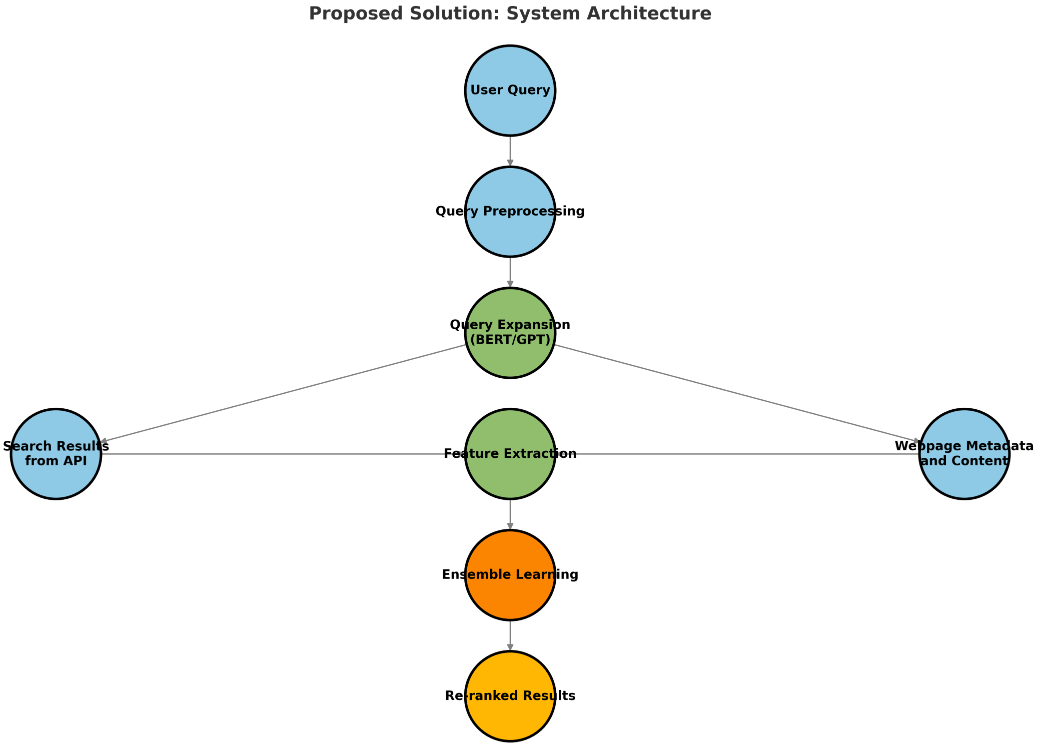

Fig. 1. System Architecture of the Proposed Solution.

where β is the tunable parameter that balances the effect of both measures.

The similarity scores are utilized as additional features in the ensemble model to enhance its predictions. The similarity scores are included as additional features in the ensemble model to strengthen its predictions.

D. Addressing Limitations of Existing Approaches

The proposed strategy can be seen to nullify some predominant drawbacks present in previous methodologies as follows:

- **Semantic Gap:** The framework captures both syntactic and semantic relationships, thereby overcoming the limitations of link-based and shallow embedding methods by applying transformer-based models influenced by JSD.
- **Scalability:** Ensemble learning leverages the principles of practical parallelism and model aggregation to process massive datasets efficiently.
- **Dynamic Adaptability:** Transformers address the static relevance model problem of traditional techniques by enriching queries based on context.
- **Integrated Robust Features:** The ensemble framework integrates diverse features, such as metadata relevance and semantic similarity, to deliver more accurate and reliable rankings.

.

Working of Proposed Solution

This framework begins by preprocessing (Algorithm 1) queries to clean, normalize, tokenize, and enrich them using transformer models like BERT or GPT. The refined query Q' is then sent to a search engine API to retrieve an initial set of $W = \{W_1, W_2, \cdots, W_p\}$ along with their metadata such as titles, snippets, and URLs.

Next, webpage metadata and content are converted into numerical features using embeddings or transformers. Multiple similarity measures, including Euclidean Distance (ED) and the Jensen-Shannon Divergence (JSD), are employed to enhance performance. ED and JSD measure the geometric distance and the probabilistic distance, respectively. Then, the weighted individual similarity scores are combined into a final similarity score that effectively balances geometric and probabilistic similarities. The aggregated similarity scores and the features of the relevant metadata are provided as input to an ensemble learning model such as Random Forests, Gradient Boosting, or Voting Classifiers to compute the relevance scores $R(W_i, Q)$ of each webpage. These relevance scores are then used to recalculate the ordering of the web pages and retrieve a more relevant list that considers the user's intent, the quality of the content, and semantic similarity. This framework, which encompasses all the components of the search ranking process, performs better than conventional methods in terms of relevance, flexibility, and expandability.

```
Algorithm 1 Proposed Solution: Framework for Webpage Retrieval
Require: Query of the user $Q$
Ensure: List of relevant webpages in rank $R = \{W_1, W_2, \ldots, W_n\}$
Step 1: Query Preprocessing
First, normalize and tokenize the query $Q$. The process includes eliminating stop words which are unnecessary
    terms followed by text standardization and tokenization of the text into its basic elements.
Step 2: Query Expansion
To expand the query, we employ a powerful transformer model based on the transformer architecture (e.g., BERT
    or GPT) to develop the query with other relevant keywords:
        $Q' = Q \cup \{k_1, k_2, \ldots, k_m\}$
    This guarantees that the search includes a broader and more relevant set of information.
Step 3: Retrieval of First Results
A list of relevant web pages $W = \{W_1, W_2, \ldots, W_p\}$ along with meta information about the web pages, including
    titles, excerpts, and URLs, should be obtained by sending the enriched query $Q'$ to a search engine API.
Step 4: Extracting Features
For each webpage $W_i$, gather essential information:
    • Gather metadata ($m_1, m_2, m_3$) and full text ($C$).
    • Convert the query $Q$ and the webpage contents into embeddings using sophisticated word representation
      models (e.g., GloVe or transformer-based embeddings).
Step 5: Semantic Similarity Computing
To identify the level of similarity between each webpage and the query, two different similarity measures are
    employed:
        $S_{\text{Euclidean}}(Q, W_i) \quad \text{and} \quad S_{\text{JSD}}(Q, W_i)$
    Both scores are combined to calculate the final similarity score:
        $S_{\text{Combined}}(Q, W_i) = \beta \cdot (1 - S_{\text{Euclidean}}(Q, W_i)) + (1 - \beta) \cdot S_{\text{JSD}}(Q, W_i)$
    where $\beta$ is a control parameter that regulates the weight given to each of the two measures.
Step 6: Ensemble Learning for Ranking Webpages
The final relevance score can be calculated by using a random forest, gradient boosting, or voting classifiers-based
    ensemble learning model.
        $R(W_i, Q) = f_{\text{ensemble}}(S_{\text{Combined}}, \text{features}(W_i))$
    This approach captures various relevance aspects to produce more accurate rankings.
Step 7: The Final Ranking is Refined
Rearrange the webpages in the order of $R(W_i, Q)$ to show the most relevant outcomes at the top.
return The final ranked list $R$, which ensures that the pages most similar to the query appear at the top.
```

TABLE I
QUERIES AFTER QUERY EXPANSION

User Query	Expanded Query (Query Expansion)
climate change 2023	impact of climate change in 2023, global warming, environmental effects, climate change trends, climate action in 2023
jaguar	jaguar animal, jaguar luxury car, jaguar sports car, jaguar species, made by jaguar
best smartphones for photography	best smartphones for photography in 2023, smartphones with best cameras, mobile cameras that are best for photography, best camera features in budget smartphones
small business loan process	small business loan application process, government small business loan, types of small business financing, veteran small business loan process
data science careers	data science job opportunities, data science career, data scientist roles and responsibilities, data scientist job description, data science career path 2023
machine learning algorithms	types of machine learning algorithms, supervised learning algorithms, unsupervised learning algorithms, new trends in machine learning, best machine learning algorithms for beginners
travel tips for Europe	The best travel tips for Europe in 2023, how to travel Europe on a budget, itineraries around Europe, Europe travel safety tips, a backpacking guide through Europe

V. Experimentation and Results

This section combines rigorous testing of the proposed framework using real-world datasets and compares its performance to that of baseline approaches. A detailed description of the experimental setup, performance evaluation, and ablation study is also included.

A. Dataset and Experimental Setup

For experimentation, the ClueWeb12-B13 dataset has been employed, one of the biggest and most reliable datasets in information retrieval, spanning over 52 million English web pages. We collected user queries from the TREC Web Track (2013-2014) to replicate actual search behaviour. The collection comprises approximately 100 queries with relevance assessments on a 0-3 scale, ensuring they are relevant to the current search scenarios.

Preprocessing Steps:

- *Query Preprocessing:* The user queries underwent tokenization, followed by normalization and BERT expansion, to produce additional contextually relevant keywords, as shown in Table I.
- *Document Preprocessing:* The system removed all HTML tags and non-essential content, including advertisements and navigation bars. The content underwent tokenization before being converted into vector embeddings by pre-trained transformers.
- *Metadata Extraction:* The system extracted and encoded additional features, including title, snippet, content length, and backlinks, as either numerical or categorical features.

Evaluation Metrics:

- *Precision@10 (P@10):* The percentage of relevant documents among the top ten.
- *Normalized Discounted Cumulative Gain (NDCG):* A measure of ranking quality considering relevance scores.
- *Mean Reciprocal Rank (MRR):* Assess how early the first relevant document appears in the ranked list.
- *F1-Score*: The harmonic mean of precision and recall as an overall retrieval quality measure.

All experiments were conducted using Python (scikit-learn, transformers) and performed on a computer with an NVIDIA RTX 3090 GPU and 64 GB of RAM.

B. Performance Evaluation

The proposed framework was compared to the following baselines, which represent several search ranking paradigms.

- *PageRank:* The original link-based ranking algorithm, now outdated, formed the basis of early web search engines, which ranked pages by their link structure.
- *Learning to Rank (LTR):* A gradient boosting machine learning based web page ranking method for optimized retrieval performance.
- *Transformer-based Ranking:* The current best standalone BERT ranker that uses a deep understanding of documents to rank them by using the current state of the field.

Results: Figure 2 presents the evaluation metrics for all methods.

The experimental results in Figure 2 demonstrate that the performance of the proposed method is better than that of the conventional ranking methods. Some key findings are:

- **Higher Ranking Accuracy**: The proposed framework outperforms the standalone transformer model, achieving 2.6% improvement in P@10, 2.6% improvement in NDCG and 7.9% improvement in MRR compared to the transformer-based ranker.
- **Improved Handling of Query Ambiguity**: The use of semantic similarity measures is crucial in enhancing retrieval relevance, particularly for ambiguous or multi-intent queries. For example, in a query like 'apple,' the query can refer to the fruit, the company, or the technology, and our method can determine the user's intent, which enhances the precision of rankings. The reason is that our method uses semantic similarity measures to help distinguish between the different meanings of a query.
- **Scaling and Time Analysis**: The proposed method strikes a suitable balance between computational time and retrieval quality, making it easily applicable to large-scale search engine contexts without significantly increasing processing time.

.

C. Ablation Study

To determine the contribution of each primary module, an ablation study was conducted by removing one component at a time to assess its impact on system performance.

- **Without Query Expansion:** The absence of query expansion resulted in a 5.2% decrease in NDCG, underscoring its importance in enhancing user intent understanding.
- **Without Semantic Similarity Features:** Excluding semantic similarity metrics (e.g., Euclidean Distance and JSD) decreased the F1-Score by 4.9%, indicating that they are helpful for contextual matching.
- **Without Ensemble Learning:** Replacing the ensemble model with a single learning algorithm, such as Gradient Boosting, resulted in a 3.7% drop in P@10, confirming that ensemble techniques significantly improve robustness and precision.

The ablation study demonstrates that query expansion, semantic similarity, and ensemble learning work effectively in combination.

VI. Result Analysis

The results confirm the framework effectively addresses existing method weaknesses. Key findings are as follows:

- **Enhanced Ranking Robustness:** Ensemble learning learns multiple feature representations and models,

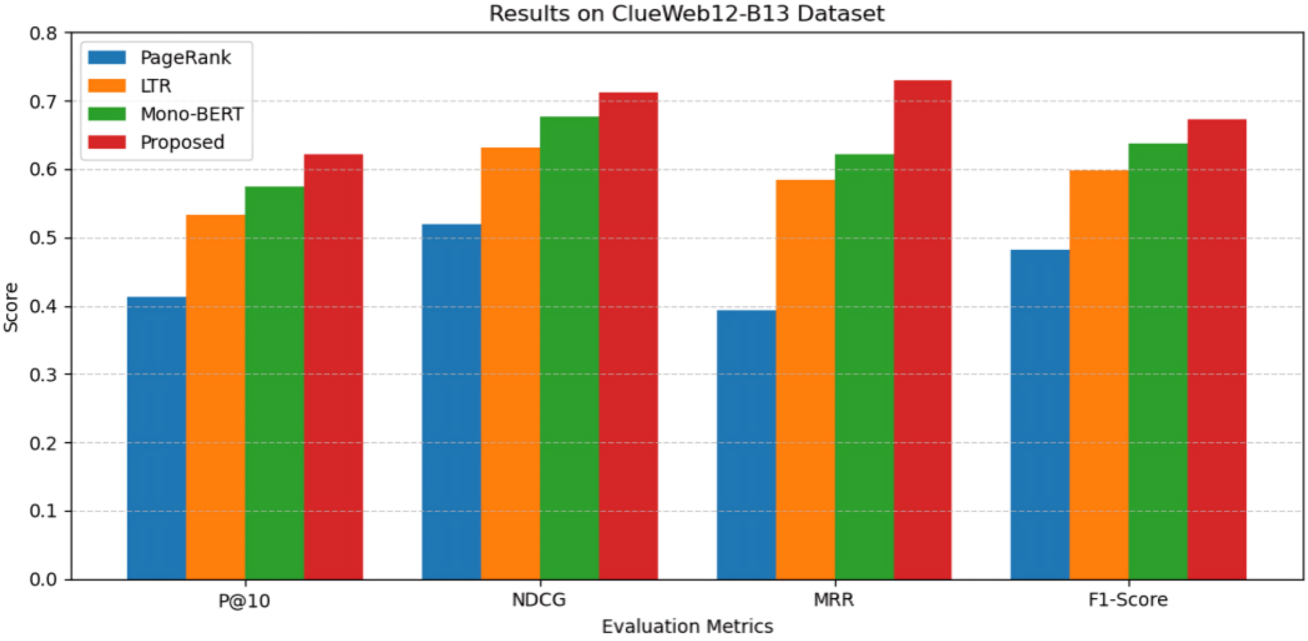

Fig. 2. Performance comparison across different retrieval methods

thereby improving ranking stability and accuracy. It strikes a balance between recall and precision, unlike standalone models.

- **Semantic Matching:** The implementation of semantic similarity metrics in ranking operations leads to better document retrieval accuracy for relevant content that matches user search intent, particularly when users enter ambiguous or complex search terms. The ranking system operates to find matches between the search query content and its context.
- **Trade-offs Between Efficiency and Accuracy:** The computational cost of transformer models brings an enhancement in semantic understanding, whilst the ensemble method is trained using precomputed features and results from ensembling simpler models.
- **Applicability to Diverse Queries:** The proposed framework demonstrated significant advantages in handling diverse query types, particularly those requiring semantic understanding and contextual awareness. For instance, in information-seeking queries like 'climate change impact in 2023,' the framework duly focused on recent scientific reports and forecasts, while traditional ranking algorithms tended to return generic or outdated results. In case of an ambiguous query like 'jaguar' which can be interpreted as an animal, a car or a football club, the system employed semantic similarity (a technique that measures the similarity between two queries based on the meaning) and query expansion (a method of expanding the original query with related terms to get more relevant documents) to resolve the ambiguity and to return contextually meaningful results. The ensemble learning technique, combined with transformer-based query expansion, retrieved specific, high-quality resources for sparse term queries with long tails, such as "small business loan application process for veterans." The framework utilized metadata relevance (price and features) to sort through pages that discussed both photography and budget clearly for multifaceted queries, such as "Best Smartphones for Photography under $500".

VII. Experimentation on MS MARCO Passage Ranking Dataset

The proposed framework is also validated through experiments conducted on the MS MARCO Passage Ranking Dataset **MS MARCO Passage Ranking Dataset https://microsoft.github.io/msmarco/**, which serves as a standard benchmark for information retrieval research. The dataset comprises actual, anonymized user queries along with web document passages, providing a practical evaluation platform for assessing retrieval performance across various large-scale, diverse query scenarios.

A. Dataset and Setup

MS MARCO (Microsoft MAchine Reading COmprehension) is composed of over 1 million annotated passages paired with natural language queries. For our evaluation, we selected a representative subset of 1000 queries and their associated candidate passages.

Preprocessing Steps:

- **Query Preprocessing:** User queries underwent normalization, followed by tokenization and expansion using BERT to include semantically meaningful and related keywords.
- **Passage Cleaning:** The system eliminated HTML tags, along with boilerplate text and scripts, that were

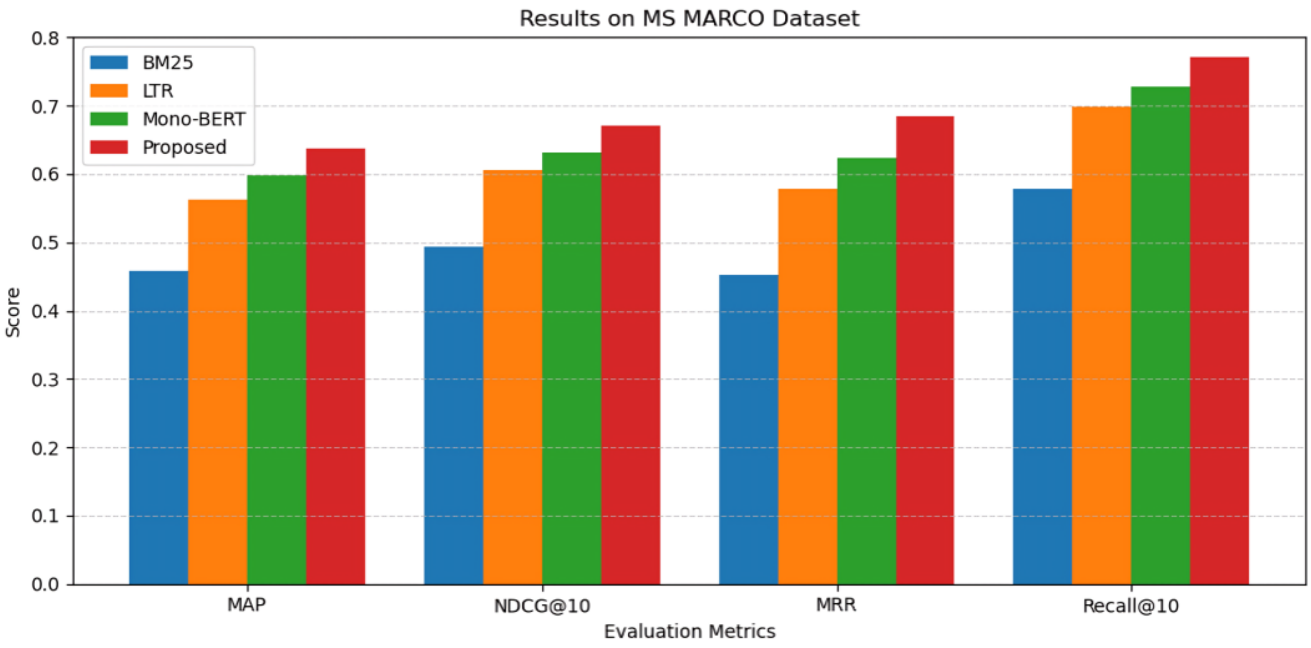

Fig. 3. Performance comparison across different retrieval methods

not relevant to the content. The cleaned passages received their embeddings through pre-trained BERT encoders.

- **Feature Extraction:** The ensemble model received passage length information, along with source domain data and content embeddings, which were properly formatted.

Evaluation Metrics:

- **Mean Average Precision (MAP)**: The system calculates the average precision value for multiple queries.
- **Normalized Discounted Cumulative Gain (NDCG@10)**: The system evaluates ranking quality through a system that considers graded relevance.
- **Mean Reciprocal Rank (MRR)**: The system measures the position of the first relevant result in the search results.
- **Recall@10**: The system evaluates the number of relevant documents found among the first ten search results.

The experiments utilized Python as the programming language, along with the scikit-learn and transformers libraries, and were run on an NVIDIA RTX A6000 GPU system with 128 GB of RAM.

B. Performance Comparison on MS MARCO Dataset

We compared the proposed framework against three strong baselines: BM25 [22], BERT-based mono-retriever [23], and Learning-to-Rank (LTR) with LambdaMART [24].

VIII. Result Analysis on MS MARCO

As shown in Figure 3 the proposed framework achieved strong performance on the MS MARCO dataset, with a MAP of 0.631 and MRR of 0.673, outperforming both traditional and neural baselines. Its high Recall@10 (0.762) reflects adequate coverage of relevant results. These gains stem from combining ensemble learning with semantic similarity, enabling precise and context-aware retrieval.

IX. Conclusion

This paper introduced a novel framework that combines ensemble learning with semantic similarity measures for improved webpage retrieval. The proposed method achieves enhanced relevance rankings, improved query understanding, and enhanced adaptability as compared to traditional methods. Hence, experimental results consistently show superior performance over baseline methods, enhancing relevance, scalability, and user satisfaction.

References

[1] Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7), 107–117.

[2] Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms* (pp. 668–677). SIAM.

[3] Xing, W., & Ghorbani, A. (2004). Weighted PageRank algorithm. In *Proceedings of the Second Annual Conference on Communication Networks and Services Research* (pp. 305–314). IEEE.

[4] Fujimura, K., Tanimoto, K., & Sugisaki, M. (2005). EigenRumor: Ranking blogs by the principal eigenvector of the comment network. In *Proceedings of the 15th International Conference on World Wide Web* (pp. 22–23).

[5] Hao, J., & Li, M. (2015). Enhanced PageRank algorithm with content relevance. *International Journal of Digital Content Technology and its Applications*, 9(5), 57–63.

[6] Kelotra, A., & Sharma, S. (2015). A time-aware ranking model for web pages. In *Proceedings of the 3rd International Symposium on Women in Computing and Informatics* (pp. 116–121).

[7] Gupta, V., & Saini, P. (2017). Automatic query expansion using semantic filtering. In *Proceedings of the IEEE International Conference on Computing, Communication and Automation* (pp. 58–63). IEEE.

[8] Azad, H. K., & Deepak, A. (2019). Query expansion techniques for information retrieval: A survey. *Information Processing & Management*, 56(5), 1698–1735.

[9] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR).*

[10] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543).

[11] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

[12] Brown, T., Mann, B., Ryder, N., & Others. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.

[13] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (pp. 5998–6008).

[14] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

[15] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232.

[16] Nogueira, R., & Cho, K. (2019). Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085.*

[17] Gao, T., Yao, X., & Chen, D. (2021). SimCSE: Simple contrastive learning of sentence embeddings. *Proceedings of the ACL 2021.*

[18] Zamani, H., Craswell, N., Taylor, M., & Smucker, M. (2022). Ensemble learning for web ranking. *Information Retrieval Journal.*

[19] Li, J., Sun, C., Wang, H., & Zhao, J. (2023). Rethinking knowledge-augmented retrieval models. *arXiv preprint arXiv:2302.12345.*

[20] Wu, X., Zhang, L., & Zhou, H. (2023). Reinforcement learning for adaptive search ranking. *NeurIPS 2023.*

[21] Chen, Y., Liu, M., & Wang, R. (2024). Multimodal retrieval techniques for intelligent search systems. *Proceedings of SIGIR 2024.*

[22] Stephen Robertson and Hugo Zaragoza, *The probabilistic relevance framework: BM25 and beyond*, Foundations and Trends in Information Retrieval, 3(4), 333–389, 2009. doi: 10.1561/1500000019.

[23] Rodrigo Nogueira and Kyunghyun Cho, *Passage Re-ranking with BERT*, arXiv preprint arXiv:1901.04085, 2019. arXiv:1901.04085.

[24] Chris Burges, *From RankNet to LambdaRank to LambdaMART: An Overview*, Microsoft Research Technical Report MSR-TR-2010-82, 2010. Link.