Enhancing Information Retrieval through Ensemble Methods and Re-ranking Strategies

Vaishnavi Shah
Computer Science
University of Massachusetts
Amherst
Amherst MA USA
vaishanvisha@umass.edu

Shivam Raj
Computer Science
University of Massachusetts
Amherst
Amherst MA USA
shivamraj@umass.edu

ABSTRACT

In the domain of information retrieval, combining ensemble retrieval models with re-ranking strategies offers a promising approach to enhance performance compared to traditional single-model methods. This research explores a range of ensemble methods, including majority vote, Borda count, and various scoring mechanisms such as average, weighted average, max, min, linear combination, and rank-based sorting. Diverse retrieval models, spanning from term-matching approaches like BM25 to neural models capturing semantic relationships, provide a comprehensive view of relevance. Additionally, the study examines re-ranking strategies, with a focus on the efficient contriever, to fine-tune search results. The goal is to create stable, robust, and personalized information retrieval systems. A thorough comparison of these ensemble and re-ranking methods offers nuanced insights into their strengths, weaknesses, and trade-offs, aiding informed decision-making for system optimization.

KEYWORDS

Ensemble methods, re ranking, flant5, language models, bert, contriever.

INTRODUCTION

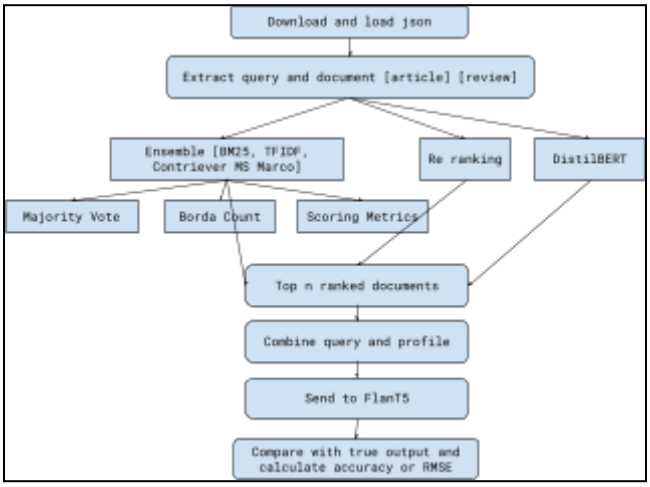
This study addresses the challenge of efficient information retrieval by exploring ensemble retrieval models and diverse re-ranking strategies to optimize performance. Various ensemble methods, including majority vote, Borda count, and scoring mechanisms, are investigated for their potential in combining the strengths of different retrieval models like BM25 and neural models. The study also explores re-ranking strategies, employing contriever and BM25. Further, DistilBERT is tested for efficiency and competitive performance to the contriever ms marco. Motivated by the goal of accurate and personalized results, the research conducts a comprehensive comparison of ensemble and re-ranking methods. Overall, this work contributes to the advancement of effective and adaptive information retrieval systems which take user profiles into consideration.

RELATED WORK

In recent literature, several notable works have contributed significantly to the field of information retrieval and evaluation frameworks. In the landscape of information retrieval, the algorithms BM25 and TF-IDF have played pivotal roles, with Robertson and Zaragoza's work introducing BM25 as an enhancement to the classic TF-IDF model [7]. The advent of Contriever, an unsupervised model introduced by Gao et al., has showcased remarkable performance, particularly in comparison to traditional methods like BM25 [6]. DistilBERT, a distilled version of BERT by Sanh et al., has emerged as a computationally efficient yet highly effective model, finding widespread adoption across various natural language processing tasks [8]. Re-ranking strategies, as explored by Huang et al. in "Learning to Rank with Neural Network for Ad Hoc Information Retrieval" [9], have demonstrated significant promise by integrating neural networks into the re-ranking process. Recently, FLANT5, a variant of FLANT, has aligned with the success of FLANT5 in few-shot learning scenarios, as highlighted in Li et al.'s paper "FLANT: Feature-wise Transformation for Few-shot Learning and Large-scale Classification" [10]. LaMP (Language Model Pre-training) offers a robust evaluation framework encompassing diverse language tasks and multiple entries for user profiles. Notably, it introduces a retrieval augmentation approach that constructs personalized prompts for large language models by retrieving personalized items from user profiles [1]. PreTTR (Precomputing Transformer Term Representations) addresses query-time latency concerns in deep transformer networks, enhancing their practicality in real-time ranking scenarios. By precomputing part of the document term representations at indexing time and merging them with query representations at query time, PreTTR achieves reduced query-time latency [2]. The Ensemble Feature Selection (EFS) approach leverages multiple Feature Selection algorithms to enhance feature identification, particularly useful for analyzing datasets and determining subsets of relevant features [3]. BERT-based text ranking models have revolutionized ad-hoc retrieval, emphasizing the importance of considering cross-document interactions and query-specific characteristics in ranking models [4]. DistilBERT, retaining 97% of BERT performance, has demonstrated its efficiency in various benchmarks such as

GLUE and BEIR [5]. Contriever, an unsupervised model, outperforms BM25 on multiple datasets for Recall@100 on the BEIR benchmark [6]. In our work, we draw inspiration from these advancements and incorporate retrieval models like DistilBERT and Contriever pre-trained with MS-MARCO into the LaMP benchmark, assessing their impact on performance.

IMPLEMENTATION



Various strategies have been employed in this comparative study.

1 Dataset Preparation:

For our research, we selected dataset 2 and dataset 3, both of which were clean and required no additional preprocessing. In dataset 2, we used the article and profiles sections for the query and document list, respectively. In dataset 3, we employed the review and profiles sections for the same purpose. The article or review and profile sections serve as candidate documents for ranking, with the query used as input for retrieval models. The profiles are stored for further analysis. The ratio determines the top n relevant documents from the total. Due to the vast number of profiles in dataset 3, random sampling was performed, fetching 50 profiles at a time to ensure manageable training in constrained time.

2 Retrieval Models:

2.1 BM25

BM25, a probabilistic information retrieval model, was applied to the dataset to retrieve profiles relevant to the given query. Its parameters were tuned for optimal performance.

2.2 TF-IDF

The Term Frequency-Inverse Document Frequency (TF-IDF) model, known for capturing term importance in documents, was utilized to retrieve profiles based on the query.

2.3 Contriever MS Marco

Contriever MS Marco, a pre-trained model designed for information retrieval tasks, was employed to enhance the retrieval process. Its specific architecture and capabilities were leveraged to cater to the nuances of the dataset.

3 Ensemble Techniques:

The ensemble approach is motivated by the understanding that these models operate on different principles and excel in different aspects of information retrieval. While BM25 and TF-IDF provide solid statistical foundations, Contriever MS Marco introduces a neural network's ability to grasp intricate semantic nuances. By combining these models, the ensemble aims to create a more comprehensive view of relevance, accommodating both statistical patterns and semantic relationships.

3.1 Majority Vote

The outputs of BM25, TF-IDF, and Contriever MS Marco were combined using a majority vote ensemble technique. All the top *n* documents receiving a majority vote were selected for subsequent analysis.

3.2 Borda Count

Borda scores were calculated for each document based on their rankings from individual retrieval models. Documents were then ranked according to their cumulative Borda scores.

$$\text{Borda Score}(D) = \sum_{i=1}^{n} (n - i) \text{ for each document } D$$

3.3 Scoring Metrics

Various scoring metrics, including average, weighted average, max, min, linear combination, sorting based on rank, and product, were employed to assign scores to each document. The documents were subsequently ranked based on each metric.

average	$= \frac{\sum_{i=1}^{n} \text{Score}_i}{n}$
weighted average	$= \frac{\sum_{i=1}^{n} \text{Weight}_i \cdot \text{Score}_i}{\sum_{i=1}^{n} \text{Weight}_i}$

sorting based on rank	$= \frac{\sum_{i=1}^{n} \text{Rank}_i}{n}$
max	$= \max(\text{Score}_1, \text{Score}_2, \ldots, \text{Score}_n)$
min	$= \min(\text{Score}_1, \text{Score}_2, \ldots, \text{Score}_n)$
linear combination	$= \sum_{i=1}^{n} \text{Weight}_i \cdot \text{Score}_i$
product	$= \prod_{i=1}^{n} \text{Score}_i$

4 **Re- ranking:**

Combined BM25, which quickly handles various queries, with Contriever MS Marco, known for its context understanding. BM25 gave us an initial ranking of documents. Followed by that, we used the top documents from this list as input for Contriever MS Marco. This neural-based model excels at understanding context and refining document relevance. The re-ranked results were sorted to fine-tune the document order, creating a more precise and contextually relevant ranking. This combination optimizes BM25's speed with Contriever MS Marco's contextual understanding for better information retrieval.

5 **DistilBERT:**

Incorporating DistilBERT into our research methodology is driven by its efficiency in capturing contextual nuances with reduced computational demands. As a distilled variant of BERT, DistilBERT maintains robust language understanding capabilities while ensuring faster processing times. Its adeptness in distilling complex contextual relationships aligns with our objective to enhance information retrieval efficiency. The inclusion of DistilBERT contributes to the optimization of our system, striking a balance between computational expediency and language comprehension for improved query-document relationships. Its ability to understand contextual relationships and generate embeddings was explored.

6 **Integration and Evaluation:**

6.1 Ranking Integration:

Based on the scores, from the ensemble techniques, scoring metrics, re-ranking, and DistilBERT achieved by the documents were sorted and integrated back into the original data. This updated list replaced the existing profiles, reflecting the refined user profiles.

6.2 Accuracy and Root Mean Squared Error (RMSE) Calculation:

The enriched query, now containing article or review and the relevant profiles, was fed into the Flant5-large model. The results were used to calculate the accuracy or RMSE of the overall system. The accuracy metric served as a quantitative measure of the effectiveness for DATASET2. RMSE was employed as a key evaluation metric to assess the predictive accuracy of the ensemble models on DATASET3.
Additionally, eval.py [1] files have also been used for evaluation purposes.

7 **Parameter Tuning:**

Systematic parameter tuning was conducted to investigate the impact of variations in parameters, such as the ratio, on the performance of the different retrieval and ranking methods. This step provided insights into the sensitivity of the system to parameter adjustments.

RESULTS

We evaluated the performance and determined whether the category was accurately categorized by using overall accuracy, RMSE as a criterion. The classification results of all models are examined and TABLE I provides a summary of the findings on dataset 2. Similarly, we evaluated the performance and determined whether the rating was correctly judged. The classification results are mentioned in TABLE 2.

DATASET 2	Ratio = 0.6	Ratio = 0.8	Ratio = 0.4
Majority votes	62.90%	59.10%	61.00%
Borda count	64.50%	62.66%	56.12%
Re ranking	67.80%	69.34%	64.20%
average	68.16%	58.40%	62.00%
Weighted average	70.40%	65.60%	66.33%
max	46.70%	59.23%	61.60%
min	68.90%	50.22%	60.30%
Linear combination	64.20%	58.45%	66.59%

product	55.84%	51.07%	58.95%
Sorting based on rank	63.70%	63.45%	67.77%
DistilBERT	73.20%	69.90%	70.01%

DATASET 3	Ratio = 0.6	Ratio = 0.4	Ratio = 0.8
Majority votes	0.5635	0.5823	0.5683
Borda count	0.9087	0.9256	0.8621
Re ranking	0.5212	0.5012	0.4990
average	0.6523	0.6235	0.6109
Weighted average	0.5377	0.6568	0.6992
max	0.6864	0.7023	0.7157
min	0.8133	0.8173	0.8223
Linear combination	0.7901	0.7389	0.7549
product	0.6220	0.6779	0.6910
Sorting based on rank	0.6432	0.6312	0.6094
DistilBERT	0.4472	0.4897	0.5023

Analyzing the results in DATASET2 and DATASET3 provides valuable insights. Majority Votes and Borda Count exhibit stability with a slight decrease in when ratio is increased, suggesting a negative impact from a larger proportion of documents in the profile which may not be relevant and has lower scores. Re-ranking consistently improves with a larger ratio, highlighting the positive influence of including more documents on refining document ranking. Scoring metrics (Average, Weighted Average, Max, Min, Linear Combination, Product, Sorting based on Rank) show lower accuracy with larger ratios in DATASET2, indicating their dependence on a richer set of 'relevant' profile documents. DistilBERT demonstrates relatively stable performance across different ratios. DATASET3 displays similar trends, underlining the significance of an extensive set of relevant documents for various methods. The impact of increasing profile documents on RMSE varies across methods. Because of time limitations and the vast amount of data, assessing the validation and testing sets was not feasible.

FUTURE SCOPE

The future scope involves in-depth exploration and optimization. Firstly, a focused investigation into the effectiveness of Reciprocal Rank (RR) as an information retrieval metric offers potential insights. Evaluating its robustness in capturing document relevance can enhance its role as an evaluation criterion.

Further research aims to refine the integration of BERT into the ensemble by exploring different variants and configurations. Fine-tuning strategies and understanding the interactions between BERT and other models will be pivotal in enhancing the ensemble's overall efficacy.

The inclusion of Relevance Model 3 (RM3) opens doors to diversifying the ensemble's components. Future work may involve integrating a broader spectrum of retrieval models, each contributing unique strengths. Exploring alternative language models beyond BERT can provide valuable insights into the impact of different language representations on information retrieval.

CONCLUSION

In conclusion, this research has undertaken a comprehensive exploration of ensemble-based information retrieval strategies, considering the integration of various retrieval models and re-ranking techniques. The experimental results provide valuable insights into the performance of ensemble models, emphasizing the benefits of combining diverse retrieval methods for enhanced accuracy and relevance in information retrieval tasks.

REFERENCES

[1] Salemi, A., Mysore, S., Bendersky, M., & Zamani, H. (2023). LaMP: When Large Language Models Meet Personalization. arXiv preprint arXiv:2304.11406.

[2] MacAvaney, Sean, et al. "Efficient document re-ranking for transformers by precomputing term representations." Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020.

[3] Mera Gaona, M. F. 2021. Selection of relevant features to support automatic detection of epileptiform events. http://repositorio.unicauca.edu.co:8080/xmlui/handle/123456789/8554.

[4] Chen, X., Hui, K., He, B., Han, X., Sun, L., & Ye, Z. (2021). Co-BERT: A Context-Aware BERT Retrieval Model Incorporating Local and Query-specific Context. arXiv preprint arXiv:2104.08523.

[5] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

[6] Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., & Grave, E. (2021). Unsupervised dense information retrieval with contrastive learning. arXiv preprint arXiv:2112.09118.

[7] Robertson, S., & Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. Now Publishers Inc.

[8] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

[9] Huang, P. S., He, X., Gao, J., Deng, L., Acero, A., & Heck, L. (2013). Learning deep structured semantic models for web search using clickthrough data. In CIKM.

[10] Li, D., Wang, Y., Hu, T., Yang, X., & Hospedales, T. M. (2021). FLANT: Feature-wise Transformation for Few-shot Learning and Large-scale Classification. arXiv preprint arXiv:2105.03856.