Ensembling Retrieval Models to Enhance User Personalization in Large Language Models

Tejaswini Amaresh*
Shrutiya Mohan*
tamaresh@umass.edu
smohan@umass.edu
University of Massachusetts Amherst
Amherst, MA, USA

ABSTRACT

This paper explores the aspect of ensembling retrieval models to craft a nuanced and adaptable approach to user personalization within the realm of large language models. As we navigate the complexities of tailoring information retrieval to individual user preferences, the exploration of ensembling techniques stands as a beacon towards a more adaptive and user-centric future in natural language processing. The paper includes details about extensive experimentation to evaluate various ensemble combinations of BM25, BERT, and HNSW. Through rigorous assessment and comparative analyses, we present the effectiveness of the proposed ensembled models in enhancing user personalization.

KEYWORDS

Ensembling, BM25, BERT, HNSW, Retrieval Model, Retrieval Augmentation, Large Language Model, Flan-T5, Zero shot learning, User Personalization

ACM Reference Format:
Tejaswini Amaresh and **Shrutiya Mohan**. 2023. Ensembling Retrieval Models to Enhance User Personalization in Large Language Models. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

As Large Language Models (LLMs) continue to dominate the forefront of information retrieval, the integration of user-specific preferences has become a pivotal challenge. Ensembling retrieval models, an approach that combines the strengths of multiple models, emerges as a promising avenue for incorporating user personalization into the fabric of these expansive language models.

This paper explores the fusion of three diverse models—BM25, HNSW [7] (Hierarchical Navigable Small World), and BERT (Bidirectional Encoder Representations from Transformers)—to form an ensemble that capitalizes on their individual strengths. By combining the statistical precision of BM25, the spatial efficiency of HNSW, and the contextual understanding of BERT, this research aims to create a comprehensive and adaptable framework for superior information retrieval.

By amalgamating the strengths of various retrieval strategies, this research aims to not only improve the relevance of information retrieval but also to create a dynamic and personalized experience for users navigating the expansive linguistic landscape.

2 RELATED WORK

In recent years, advancements in large language models (LLMs) have significantly enhanced natural language processing capabilities. Incorporating user personalization within these models has garnered attention due to its potential to improve information retrieval and recommendation systems. In this section, we review the existing literature on ensembling retrieval models and user personalization in large language models. Our work builds upon the foundations established by the paper [11].

Ensembling retrieval models have been widely studied in the field of information retrieval. Ensemble methods combine multiple retrieval models to improve the overall performance. This approach has proven to be effective in various retrieval tasks, such as document ranking and question answering.

In the paper [12], the authors propose a generalized ensemble model (gEnM) for document ranking. The model formulates an optimization program to obtain the optimal linear combination of these models by maximizing the mean average precision. The paper [9] describes a re-implementation of BERT, a neural model pre-trained on a language modeling task, for query-based passage re-ranking. The system achieved state-of-the-art results on the TREC-CAR dataset and the MS MARCO passage retrieval task.

The study [6] experiments the use of ranker, whose results are then fed into summarization and synthesis models, enhancing personalised results. The paper [11] proposes a method that leverages user-specific information and preferences to generate personalized responses. Their approach involves fine-tuning the language model with personalized data, such as user feedback and past interactions. Building on this architecture, the model described in this paper adopts an ensembling approach to combine the results of sparse

*Both authors contributed equally to this research.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

and dense retrieval models and further re-rank to obtain more personalized results.

The paper [10] investigates the task of full-rank retrieval of responses for dialogues. The authors compare supervised and unsupervised, dense and sparse retrieval models for this task. They explore dialogue context and response expansion techniques for sparse retrieval, as well as zero-shot and fine-tuned dense retrieval approaches. Dense retrieval models with intermediate training followed by fine-tuning perform best, and harder negative sampling techniques lead to worse effectiveness.

3 ARCHITECTURE

The end-to-end architecture behind personalized LLM output involves the following aspects:
1. Query generation function
2. Retrieval model that incorporates related user profiles for the query.
3. Retrieval augmented output from the retrieval model is formulated as a personalized prompt for the LLM.
4. The Personalized Prompt input for the LLM results in a personalized output.
Our goal is to achieve improved personalization by integrating ensemble modeling into the Retrieval model component of the architecture.

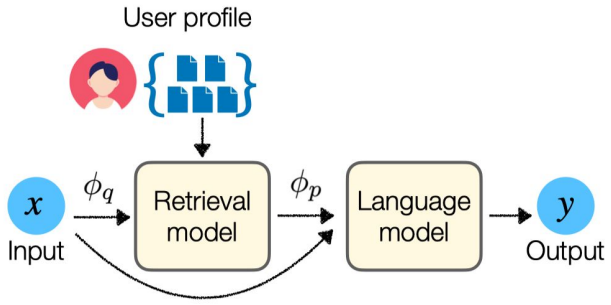

Figure 1: An overview of the retrieval-augmented method for personalizing LLMs. Φ q and Φ p represent query and prompt construction functions

4 DATASET

The LaMP benchmark is used to evaluate the effectiveness of language models in generating personalized results by utilizing information specific to individual users. We chose the 'Personalized News Categorization' [8] dataset and the 'Personalized News Headline Generation' [1] dataset. The 'Personalized News Headline Generation' dataset had the lowest scores in the evaluation section of the LaMP paper, prompting our interest in improving the metrics for this particular dataset. We intend to explore datasets related to both classification and text generation, examining how the incorporation of ensembling impacts the integration of user personalization in Large Language Models (LLMs).

4.1 Personalized News Categorization

The dataset 'Personalized News Categorization' is designed for multi-class classification, with the goal of evaluating a language model's ability to categorize news articles authored by a user (journalist) u. We employ the User-based separation in our approach, focusing on user distinctions rather than modeling the temporal aspects of the data.

4.2 Personalized News Headline Generation

The 'Personalized News Headline Generation' dataset is designed to assess the language model's proficiency in generating headlines for news articles authored by a user (u). This evaluation task gauges the model's capacity to create informative and personalized headlines, taking cues from the user's profile. In our methodology, we utilize User-based separation, concentrating on user distinctions rather than modeling the temporal aspects of the data.

5 EXPERIMENT

5.1 Experimental Setup

We integrate an ensemble approach that combines various configurations of three models—BM25, BERT, and HNSW—using weighted voting. In addition to these models, we employ a fine-tuned Language Model (LLM)[4]. Given the computational demands of the BERT model, we conduct training and validation on 100 instances of each dataset. To account for the distinctive features of sparse retrieval, dense retrieval, and contextual understanding offered by BM25, HNSW, and BERT, we assign equal weights to all ensembled combinations. Each experiment is designed to retrieve k=1 relevant document for every query. We employ k = 1 for simplicity. Increasing k lead to longer Aggregated Input Prompts (AIP) and added computational complexity in training fine-tuned LLMs.

BM25 model : We use Okapi BM25 model, which is a probabilistic sparse retrieval model. It's a refined version of the tf-idf model with document length normalization (b=0.75) and addresses document term saturation (k=1.5) .

BERT : We use zero shot BERT [3] base uncased that generates contextual embeddings for each token in a sequence. These embeddings capture the contextual intricacies and associations among words, making the model well-suited for the retrieval task of fetching user profile documents related to a given input query.

HNSW: HNSW is a dense retrieval model that is well-suited for indexing large datasets in high-dimensional spaces efficiently. It enables quick retrieval of points that are likely to be close to a given query point. We use ef (Entry Point Factor) = 50 and M (maximum number of neighbors) = 20.

Ensembling Combinations (BM25 + BERT + HNSW, BM25 + BERT, BERT + HNSW, BM25 + HNSW): To ensemble different combinations of models by weighted voting (with equal weights), the following steps are performed.
1. Get the (input, user profile) similarities scores for every user

profile of every instance for all the 3 models.
2. Get the sum of weighted scores of the ensembled combinations.
3. Get the top k (k=1) user profile for every input instance and form the Augmented Input Prompt(AIP) based on the ensembled model scores.

BERT as a re-ranker: To reduce the processing time and computational complexity of BERT and to model the output for more instances of the dataset, we perform the following steps:
1. Get (input,user profile similarity) scores for each instance of the dataset for BM25 and HNSW models.
2. Ensemble both the models with equal weight.
3. Based on the scores of the ensembled model, get the top 5 relevant user profiles for every instance.
4. Use BERT to re-rank the top 5 user profiles for every instance.

Flan-T5 Base: Flan-T5 [2] is an enhanced version of T5 that has been fine-tuned in a mixture of tasks, thereby improving the effectiveness of the zero-shot learning. The model explores instruction fine-tuning with a particular focus on (1) scaling the number of tasks, (2) scaling the model size, and (3) fine-tuning on chain-of-thought data. This model is further fine-tuned for each dataset with the most relevant documents concatenated along with the query as inputs.
1. Load and Pre-process the dataset by identifying the maximum length of the documents and perform padding if necessary.
2. Load the pre-trained model and set the parameters required for the training. For experimental purposes, the number of epochs was set to 2.
3. Train the model and upload it to HuggingFace[5].
4. Evaluate the modle on the validation data provided.

6 RESULTS

Based on the results from Table 1, we can observe that for the 'Personalized News Categorization' dataset, BM25+ HNSW re-ranked with BERT has the best accuracy and macro averaged F-1 score. But, the factor of higher number of data instances being considered cannot be ignored. Comparing the ensembled combinations that have 100 data instances, we can observe that BM25 + HNSW ensembled with equal weights of 0.5 has the best accuracy and macro averaged F-1 scores.

From Table 2, we can observe that, for the text generation dataset of 'Personalized News Headline Generation' , the BM25+HNSW scores for 4000 train instances and 1925 validation instances reranked with BERT are fairly close to BM25+ BERT scores for 100 data instances. Although BM25 + HNSW scores re-ranked with BERT have the highest ROUGE-1 and ROUGE-L scores, there is a high probability that BM25 model ensembled with BERT for more instances would result in higher ROGUE- 1 and ROGUE-L scores leading to enhancement in incorporating user personalization of Large Language Models

Overall, we can conclude that ensembled combination of BM25 and HNSW model has a high probability of enhancing accuracy in the "Personalized News Categorization" dataset. The ensembled combination of BM25 and BERT has a high probability of having enhanced ROGUE scores for the "Personalized News Headline Generation" dataset.

7 FUTURE WORK

Future research can include the ensembling of these retrieval models across a larger number of dataset instances, leveraging GPU acceleration for computational efficiency. Furthermore, there is an opportunity to improve performance by applying fine-tuning techniques to the pre-trained BERT model. Another avenue worth exploring is the training of neural retrieval models instead of relying on Zero Shot learning.

Additionally, investigating voting schemes with unequal weights for each model and observing the output of the Large Language Model (LLM) for each combination could provide valuable insights. Researching alternative ensembling methods, such as stacking, temporal ensembling, and Rank Fusion, also holds promise for enhancing the overall effectiveness of the retrieval system.

ACKNOWLEDGMENTS

To Professor Hamed Zamani and Teaching Assistant Yen-Chieh Lien for providing guidance throughout the research.

REFERENCES

[1] Xiang Ao, Xiting Wang, Ling Luo, Ying Qiao, Qing He, and Xing Xie. 2021. PENS: A Dataset and Generic Framework for Personalized News Headline Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 82–92. https://doi.org/10.18653/v1/2021.acl-long.7
[2] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416* (2022).
[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]
[4] huggingface.co. 2022. Transformers Model: T5. Retrieved June 24, 2022 from https://huggingface.co/docs/transformers/model_doc/t5
[5] huggingface.co. 2023. Fine-Tuned Models. Retrieved April 27, 2023 from https://huggingface.co/models?sort=trending&search=shrutiya%2F?fw=pt
[6] Cheng Li, Mingyang Zhang, Qiaozhu Mei, Yaqing Wang, Spurthi Amba Hombaiah, Yi Liang, and Michael Bendersky. 2023. Teach LLMs to Personalize–An Approach inspired by Writing Education. *arXiv preprint arXiv:2308.07968* (2023). https://arxiv.org/pdf/2308.07968
[7] Yu A Malkov and Dmitry A Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence* 42, 4 (2018), 824–836. https://arxiv.org/pdf/1603.09320
[8] Rishabh Misra. 2022. News Category Dataset. arXiv:2209.11429 [cs.CL]
[9] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085* (2019). https://arxiv.org/abs/1901.04085
[10] Gustavo Penha and Claudia Hauff. 2022. Sparse and Dense Approaches for the Full-rank Retrieval of Responses for Dialogues. *arXiv preprint arXiv:2204.10558* (2022). https://arxiv.org/abs/2204.10558
[11] Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. LaMP: When Large Language Models Meet Personalization. arXiv:2304.11406 [cs.CL]
[12] In-Chan Choi Wang, Yanshan and Hongfang Liu. 2015. Generalized ensemble model for document ranking in information retrieval. (2015). https://arxiv.org/pdf/1507.08586

		Fine-Tuned Flan T5-Base			**Metrics**	
Retrieval Model	*Re-Ranking Model*	*Trained Instances*	*Validated Instances*	*k*	*Accuracy*	*Macro-Avg F1*
BM25 + BERT + HNSW	-	100	100	1	0.3960	0.1304
BM25 + BERT	-	100	100	1	0.4554	0.1554
HNSW + BERT	-	100	100	1	0.3960	0.1715
BM25 + HNSW	-	100	100	1	0.4653	0.1757
BM25+HNSW	BERT	5914	1052	3	0.7966	0.4923

Table 1: Experimental Results for Lamp2U: Personalized News Categorization

		Fine-Tuned Flan T5-Base			**Metrics**	
Retrieval Model	*Re-Ranking Model*	*Trained Instances*	*Validated Instances*	*k*	*Rouge-1*	*Rouge-L*
BM25 + BERT + HNSW	-	100	100	1	0.1537	0.1363
BM25 + BERT	-	100	100	1	0.1635	0.1438
HNSW + BERT	-	100	100	1	0.1434	0.1244
BM25 + HNSW	-	100	100	1	0.1441	0.1316
BM25 + HNSW	BERT	4000	1925	3	0.1661	0.1542

Table 2: Experimental Results for Lamp4U: Personalized News Headline Generation