Enhancing User Personalization through Large Language Models: A Similar Users Approach

Siddarth Suresh
University of Massachusetts Amherst
Amherst
Massachusetts
siddarthsure@umass.edu

Tanvi Gupta
University of Massachusetts Amherst
Amherst
Massachusetts
tanvigupta@umass.edu

ABSTRACT

This research introduces a novel methodology for achieving personalized product ratings and personalized tweet paraphrasing by making use of the LaMP benchmark. The project explores a novel approach to user personalization by harnessing the power of large language models. Existing methods rely solely on individual user profiles to improve personalization. Our methodology focuses on augmentation of the user profile with profiles of similar users. We identify and analyze patterns in user behavior, aiming to uncover latent connections between similar users. Our experiment aims to answer whether aggregating the current user's data with the data of the users that are similar improves performance. Our experiment also aims to answer whether text summarizing users' profiles before being added to the input improves the ability of the large language model to classify as well as paraphrase textual content.

Keywords

Personalization; Large Language Models; Text Similarity; LaMP Benchmark;Text Summarization; ; Contriever;

INTRODUCTION

In the ever-growing field of Natural Language Processing, the demand for personalized user experiences have continually increased. . To address this need, our research introduces a methodology that harnesses the collective intelligence of similar users to augment user personalization. By adopting a proactive stance, we compute and store a comprehensive list of users exhibiting patterns similar to the particular user. The aggregation of data from both the user in focus and their analogous counterparts becomes the cornerstone of our personalized content curating strategy.

In the subsequent stages, our approach strategically selects the top k results from this amalgamated dataset. Employing summarization techniques, these results are distilled into representations, ensuring the retention of critical keywords and context. The generated prompt, born out of these summarized results, is then fed into a Large Language Model (LLM).

By integrating this prompt into the LLM, our method orchestrates the generation of paraphrased text and classification that seamlessly blends the unique context of the current user with the collective insights from similar users. The convergence of these elements yields a final output that improves traditional personalization paradigms, offering users a more nuanced, contextually rich, and engaging content experience. Through this innovative approach, we contribute to the discourse on user personalization strategies and underscore the potential of amalgamating similar user data with advanced language models.

This research uses the architectural framework proposed by LaMP which comprises of three primary stages – the generation of a query function(φq), a retrieval model (R) and a prompt construction function (φp). Using these three stages, LaMP aims to transform user related data to a all inclusive prompt for the large language model, thereby enabling large language model to classify and paraphrase text..

The research focuses on two datasets provided by the LaMP benchmark 1) Personalized Tweet Paraphrasing and 2) Personalized Product Rating. The two datasets host a plethora of user based information. There are two settings available for the datasets out of which our research focuses on the user based setting. This setting enables us to study the enhancement of personalization of unseen users.

In this project, we have attempted to answer the following two questions:

1. Does expanding a single user's profile with other similar users improve LLM performance?
2. Does summarising user's profiles to retain essential information improve LLM performance?

RELATED WORK

The landscape of personalized content recommendations has seen significant research and development in recent years. Our approach draws inspiration and builds upon existing works in user personalization, language models, and data summarization.

User Personalization with Language Models:

Prior research has explored the integration of language models for user personalization. The work by Li et al. [1] delves into the use of language models to predict user preferences based on historical data. Similarly, Cheng et al. [2] propose a collaborative filtering method using neural networks to enhance personalized recommendations. Our methodology extends these concepts by focusing on user similarity, emphasizing the collective

intelligence of users with similar behavioral patterns rather than individual preferences alone.

User Similarity and Collaborative Filtering:

Collaborative filtering techniques, as discussed by Herlocker et al. [3], have been widely employed in recommender systems. Furthermore, Koren et al. [4] introduce matrix factorization as an effective collaborative filtering approach. Our work aligns with the collaborative filtering paradigm by identifying and analyzing patterns in user behavior to uncover latent connections between similar users. This approach aims to enhance the performance of personalized content recommendations by leveraging the wisdom of the crowd.

User Data Summarization:

The concept of user data summarization has been explored in the context of various natural language processing tasks. Liu et al. [5] proposed a method for summarizing user reviews to improve sentiment analysis. Additionally, Erkan and Radev [6] discuss extractive summarization techniques for generating concise summaries from large text documents. In our study, we extend the idea of user data summarization to enhance language model performance. We investigate whether summarizing user profiles before inputting them into a language model contributes to improved performance.

Methodology

The adopted research methodology focuses on the augmentation of data pertaining to the user of interest by incorporating information from users with similar profiles. This process involves the utilization of pre-trained transformer models, specifically fine-tuned to address the task of text similarity unique to our research domain.

In detail, the methodology initiates with the generation of a comprehensive list of users who exhibit profiles akin to that of the target user. This is achieved through the application of the transformer model such as Bart or Flan T5 specifically fine-tuned on the task of text summarization. The subsequent step involves the retrieval of the top-k users with profiles most closely aligned to the input query, using the contriever transformer model.

To circumvent challenges associated with the limited input size of large language models and simultaneously preserve the essential information, a strategic summarization approach is implemented. Leveraging a transformer model fine-tuned explicitly for summarization tasks, we distill the selected user profiles into concise representations.

These condensed profiles are then employed as prompts for the Large Language Model (LLM). This strategic use of summarization serves the dual purpose of mitigating input size constraints and enhancing the LLM's capacity to make refined classification and paraphrasing decisions. This multi-stage process ensures a judicious augmentation of user data, contributing to heightened robustness and efficacy in the outcomes of our research endeavor.

THE LaMP BENCHMARK

The LaMP benchmark aims at assessing the efficacy of language models in producing personalized outputs based on user-specific information for seven diverse tasks. Of these tasks, we have selected the following two:

Personalized Text Classification

(1) Personalized Product Rating:

This task evaluates the language model's ability to predict the rating that a user u has given to a product based on the review written by u for the product. The goal is to predict an ordinal score with a range from 1 to 5, and exclusively consisting of integer values.

Personalized Text Generation

(2) Personalized Tweet Paraphrasing:

The proposed task aims to evaluate the language model's capacity to generate a paraphrased version of a tweet, considering the writing style of the user. Specifically, the language model is presented with a tweet and instructed to generate a corresponding paraphrase that captures the essence of the original tweet while aligning with the writing style of the user.

This benchmark creates each dataset in two different settings: 1) user-based separation and 2) time-based separation.

User-based Separation. We have chosen the user-based separation of the datasets for both of our tasks. In the context of user based separation, individuals are partitioned into training, validation, and test sets. Notably, the temporal aspect is disregarded in this scenario, whereby the selection of profile items, inputs, and outputs is randomized rather than time-dependent. As stated in the LaMP benchmark, To achieve personalization for a given sample (xi, yi) associated with user u, three primary components are employed: (1) a query generation function φq that transforms the input xi into a query q for retrieving from the user u's profile, (2) a retrieval model R(q, Pu, k) that accepts a query q, a user profile Pu and retrieves k most pertinent entries from the user profile, and (3) a prompt construction function φp that assembles a prompt for user u based on input xi and the retrieved entries. Consequently, the input $\bar{x}i$ for the language model is derived using the following formulation:

$$\bar{x}i = \varphi p(xi, R(\varphi q(xi), Pu, k))$$

GENERATION OF LIST OF SIMILAR USER PROFILES

In the process of constructing a comprehensive collection of similar user profiles, our methodology initiates by meticulously consolidating the entirety of each user's profile data. This systematic aggregation ensures a thorough representation of the diverse facets that constitute individual digital identities.

For the discernment and quantification of user similarity, we employ a sophisticated sentence transformer model, exemplified by the "all-MiniLM-L6-v2." This model serves as a pivotal tool in generating a finely ranked list of user profiles, grounded in their intrinsic resemblances to a specified user of interest.

The resultant catalog of ranked user profiles is conscientiously archived in an offline repository. This deliberate offline storage strategy is adopted with the aim of optimizing efficiency in subsequent data retrieval tasks while proactively mitigating potential challenges associated with real-time processing.

DOCUMENT RETRIEVAL USING CONTRIEVER

Within the scope of this project, the Contriever transformer model has been instrumental in retrieving the k most relevant user profiles based on a given input query. The functionality of the Contriever transformer model is rooted in the generation of sentence embeddings for each input query. These embeddings serve as numerical representations encapsulating the semantic meaning of the input sentences. The model then employs a dot product calculation on these embeddings, establishing a measure of similarity between sentences within the input context.

This methodology ensures a nuanced exploration of semantic relationships, allowing the Contriever transformer model to discern and prioritize user profiles that exhibit the highest degree of relevance to the provided query. By harnessing the power of sentence embeddings and similarity metrics, we aim to enhance the precision and effectiveness of user profile retrieval within the context of our research project.

PROFILE SUMMARIZATION

Given the input limitations of the Large Language Model (LLM), it becomes imperative to condense the size of each user profile retrieved using the Contriever while retaining its hidden features and essence. To accomplish this, we employ pre-trained transformers that are fine-tuned explicitly for the task of text summarization, such as 'Bart' and 'Flan t5.' These models enable us to distill the key information from each profile effectively.

Upon summarization, the condensed profiles are then forwarded to the prompt generation stage. This strategic transition ensures that the essential information, initially embedded within the user profiles, is maintained in a more concise format suitable for input to the LLM.

By integrating these summarization techniques into our methodology, we aim not only to overcome input size constraints but also to enhance the efficiency and effectiveness of the LLM by providing it with succinct yet informative prompts derived from the distilled essence of user profiles. This step contributes to the overall robustness and interpretability of our research outcomes.

PROMPT GENERATION

The Aggregated Input Prompt (AIP) is created by combining the original input query with summaries of the top-k similar user profiles. This step aims to generate a prompt that incorporates both the current user's context and insights from similar users.

Formats for input prompts for LLM:

1. For personalised product rating: concat([PPEP(P1), ..., PPEP(Pn)], ", and "). [INPUT]
2. For tweet paraphrasing: concat([PPEP(P1), ..., PPEP(Pn)], ", and ") are written by a person. Following the given patterns [INPUT]

concat is a function that concatenates the strings in its first argument by placing the string in the second argument between them. [INPUT] is the input query.

To enhance the AIP generation process, we made use of the Contriever retrieval method. This method utilizes a pre-trained Contriever model to retrieve relevant profiles based on the input query. The retrieved profiles are then integrated into the AIP which is then fed into the LLM.

Experimental Setup

We have made use of Contriever to retrieve the k best user profiles relevant to the input query. In order to generate a summarization of the user profiles, we have used the summarization pipeline from the transformers library. To fine tune the Flan-T5 model, we leverage the adafactor optimizer with a learning rate of 3×10^{-4}. We also incorporate a weight decay of 10^{-2} to prevent overfitting during training. We train our generation model for 5 epochs and use Flan-T5-base model for our experiment.

Results:

We performed the following four experiments on our data:

1. Evaluation of LLM performance with aggregating similar users' data and with summarization of the input prompt.
2. Evaluation of LLM performance without aggregating similar users' data and without summarization of the input prompt.
3. Evaluation of LLM performance with aggregating similar users' data and without summarization of the input prompt.
4. Evaluation of LLM performance without aggregating similar users' data and with summarization of the input prompt.

On performing our experiments, the following results were observed:

Tweet Paraphrasing:

	Rouge-1	Rouge-l
Prompt generated with aggregating similar users' data and with summarization of the input prompt.	0.434	0.043
Prompt generated without aggregating similar users' data and without summarization of the input prompt.	0.402	0.036
Prompt generated with aggregating similar users' data and without summarization of the input prompt.	0.396	0.035
Prompt generated without aggregating similar users' data and with summarization of the input prompt.	0.446	0.039

Personalised Product Rating:

	MAE	RMSE
Prompt generated with aggregating similar users' data and with summarization of the input prompt.	0.633	0.912
Prompt generated without aggregating similar users' data and without summarization of the input prompt.	0.533	1.0
Prompt generated with aggregating similar users' data and without summarization of the input prompt.	0.63	1.08
Prompt generated without aggregating similar users' data and with summarization of the input prompt.	0.433	0.707

From the above, metrics, it can be seen that our approach of aggregating data from similar users and generating an efficient summary out of the user profiles succeeds in performing better than using raw data to achieve user personalization.

In the context of tweet paraphrasing, the Rouge-1 and Rouge-l scores serve as key evaluation metrics for assessing the quality of paraphrased content. Notably, when prompts are generated by aggregating data from similar users and incorporating summarization, the Rouge-1 score reaches 0.434, outperforming alternative scenarios. This indicates a superior ability to capture content overlap between the original and paraphrased tweets. Similarly, the Rouge-l score attains 0.043, demonstrating a commendable level of linguistic similarity.

In the domain of personalized product rating, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are crucial metrics to gauge the accuracy of predicted product ratings. Here, the approach of aggregating similar users' data and summarizing input prompts outshines other strategies. The MAE of 0.633 and RMSE of 0.912 in this scenario showcase a superior ability to predict user-specific product ratings compared to alternative methods.

Overall, the consistent trend across both tasks reinforces the efficacy of our approach. By leveraging aggregated data from similar users and generating concise summaries, our method excels in achieving enhanced user personalization. These results underscore the importance of considering user similarity and efficient summarization techniques in the pursuit of optimizing language model performance for diverse tasks.

CONCLUSION

In conclusion, this research introduces a novel approach to user personalization by leveraging large language models and the concept of user similarity. The methodology explores the aggregation of data from both the current user and similar users, leading to a more nuanced and contextually rich understanding of user needs. The integration of summarization techniques further refines the input for Large Language Models (LLMs), resulting in personalized outputs that blend the unique context of the current user with collective insights from similar users.

The LaMP benchmark provides a comprehensive evaluation framework, focusing on tasks such as Personalized Text Classification and Personalized Text Generation. Our experiments, particularly in the domain of Personalized Tweet

Paraphrasing, showcase the effectiveness of the proposed approach. The incorporation of Contriever for profile retrieval and the Flan-T5 model for summarization contribute to the success of the personalized content generation process.

FUTURE WORK

Despite the promising results, there are avenues for future exploration and improvement. One potential direction is to enhance the methodology's scalability, especially when dealing with a large number of users and diverse datasets. Additionally, the research could delve into more advanced summarization techniques and model architectures to further improve the quality of generated prompts.

Moreover, investigating the impact of incorporating additional user-specific features or contextual information could enhance the personalization capabilities of the approach. The exploration of dynamic user profiles that adapt to changing preferences over time could also be a fruitful avenue for future research.

Furthermore, the generalization of the proposed method to different domains and datasets would provide insights into its robustness and adaptability. Collaborations with industry partners or real-world applications could facilitate the deployment and validation of the approach in practical scenarios.

In summary, the presented research lays the foundation for a personalized content recommendation system that harnesses the collective intelligence of similar users. Future work can build upon this foundation to refine and extend the methodology for even more effective and widely applicable user personalization strategies.

REFERENCES

[1] Li, W., Lu, J., Wu, Z., & Yang, Y. (2015). Recommender system application developments: A survey. Decision Support Systems, 74, 12-32.

[2] Cheng, H., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., ... & Haque, Z. (2016). Wide & deep learning for recommender systems. In Proceedings of the 1st Workshop on Deep Learning for Recommender Systems (pp. 7-10).

[3] Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems (TOIS), 22(1), 5-53.

[4] Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. Computer, 42(8), 30-37.

[5] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized BERT approach. arXiv preprint arXiv:1907.11692.

[6] Erkan, G., & Radev, D. R. (2004). LexPageRank: Prestige in multi-document text summarization. In Proceedings of the 2004 conference on empirical methods in natural language processing (pp. 365-371).

[7] Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., & Wei, J. (2023). Scaling Instruction-Finetuned Language Models. arXiv preprint

Richardson, C., Zhang, Y., Gillespie, K., Kar, S., Singh, A., Raeesy, Z., Khan, O. Z., & Sethy, A. (Year). Integrating Summarization and Retrieval for Enhanced Personalization via Large Language Models. Journal/Conference Name, Volume(Issue), Page Range.