MMR-Diversified BM25 for Personalized Language Models Optimization

Harshith Reddy Takkala
Manning CICS
University of Massachusetts
Amherst
+14132306473
htakkala@umass.edu

Keerthy Kaushik Dasoju
Manning CICS
University of Massachusetts
Amherst
+14138298780
kdasoju@umass.edu

ABSTRACT

This research investigates the application of Maximal Marginal Relevance (MMR) for diversifying Best Matching 25 (BM25) results in the context of Personalized Language Models (LLMs) optimization. The motivation for this study stems from the ubiquitous challenge of information overload in contemporary information retrieval systems. Users increasingly demand more personalized and relevant content from search engines. The integration of MMR with BM25 presents an opportunity to strike a balance between relevance and diversity in search outcomes, addressing the limitations of traditional ranking methods.

Our research employs three distinct datasets—Personalized Citation Identification, Personalized News Categorization, and Personalized Scholarly Title Generation—each contributing to the exploration of MMR-BM25 synergy in diverse information retrieval tasks. The methodology involves data preprocessing using a subset of the LaMP benchmark's validation dataset, leveraging GPT-3.5 Turbo as the Large Language Model without fine-tuning. We utilize 100 random data points due to resource constraints, applying BM25 to obtain initial relevance scores. Subsequently, we extract top results, refine the query by appending a mix of document titles, articles, abstract and feed this into the LLM. The process is repeated with MMR diversification, introducing three lambda values (50.4, 0.5 and 0.6) to evaluate its impact. Our findings reveal mixed results across the datasets, with one dataset demonstrating substantial improvement. The exploration of lambda values and top-k selections in BM25 and MMR diversification is presented visually and quantitatively. The evaluation metrics include accuracy and Rouge scores. The outcomes offer insights into the nuanced relationship between BM25, MMR, and personalized LLMs, paving the way for enhanced search experiences studies.

Keywords

Maximal Marginal Relevance (MMR), Best Matching 25 (BM25), Personalized Language Models (LLMs), GPT-3.5 Turbo, Diversification, LaMP Benchmark.

1. INTRODUCTION

The rapid proliferation of digital information has led to an era where users demand highly personalized and contextually relevant content from search engines. Addressing this challenge is paramount for optimizing search experiences and mitigating the effects of information overload. In response, this research explores a novel paradigm for Personalized Language Models (LLMs) by investigating the synergistic interplay between Maximal Marginal Relevance (MMR) and Best Matching 25 (BM25) algorithms.

The primary objective of this study is to offer a nuanced solution to the limitations of traditional ranking methods by enhancing the relevance and diversity of search outcomes through the integration of MMR and BM25. By doing so, we aim to contribute to the optimization of Personalized Language Models, catering to the evolving expectations of users in contemporary information retrieval systems. Our findings reveal a mixed landscape across three distinct datasets—Personalized Citation Identification, Personalized News Categorization, and Personalized Scholarly Title Generation. Notably, personalization news categorization user-based dataset demonstrates substantial improvement, underscoring the potential of the proposed MMR-BM25 synergy. The exploration of lambda values and top-k selections in BM25 and MMR diversification is presented visually and quantitatively, providing insights into the performance of the proposed approach.

This paper unfolds in a structured manner to present a comprehensive understanding of our research. Section II provides a literature review, contextualizing our study within existing research. Section III details the methodology, outlining the data extraction, random data points selection, and the process of updating queries with BM25 and MMR results. In Section IV, we discuss conclusions in Section V. Section VI ends by referring the papers used during our research.

2. LITERATURE SURVEY

[1] The seamless integration of Maximal Marginal Relevance (MMR) not only mitigated redundancy challenges in my keyphrase extraction project but also fostered a more robust approach to information representation. By implementing MMR's diversity-based reranking methodology, the resulting keyphrase set achieved a heightened level of coherence and informativeness. The adaptability of MMR in balancing query relevance and novelty emerged as a crucial asset, offering a sophisticated solution to the intricate problem of optimizing limited display space. Through this implementation, the practical manifestation of insights gleaned from the influential paper not only enhanced the quality of keyphrase extraction but also underscored the profound impact of cutting-edge research on addressing nuanced real-world information challenges.

[2] The introduction of the LaMP benchmark in the paper "LaMP: When Large Language Models Meet Personalization" significantly impacts my research project by addressing a critical void in Natural Language Processing benchmarks. This benchmark offers a diverse evaluation framework, encompassing personalized text classification and generation tasks, aligning closely with my exploration of Maximal Marginal Relevance

(MMR) in personalized language models. The incorporation of retrieval augmentation strategies provides valuable insights into optimizing the integration of MMR and Best Matching 25 (BM25) for enhancing relevance and diversity in generated responses. LaMP's comprehensive tasks and methodologies establish it as an indispensable resource, guiding the assessment of personalized language understanding and generation in large language models.

[3] This pivotal paper on Maximal Marginal Relevance (MMR) has been instrumental in shaping and advancing my research project, which focuses on enhancing Personalized Large Language Models (LLMs) through the integration of MMR with the Best Matching 25 (BM25) algorithm. By exploring MMR's efficacy in reducing redundancy while maintaining query relevance, especially in multi-document summarization, the paper provided crucial insights applicable to my diverse information retrieval tasks. The foundational framework presented, emphasizing the balance between relevance and diversity, served as a guiding principle in optimizing query results across three distinct datasets. The paper's influence is evident in the iterative and data-driven approach adopted, directly impacting the design and implementation of MMR within the context of Personalized Language Models. Overall, this paper significantly enriched the theoretical depth and practical applicability of my research.

[4] The paper "Evaluation of Diversification Techniques for Legal Information Retrieval" significantly informs my project by addressing the challenge of information overload in legal information retrieval. Its exploration of diverse diversification methods, including MMR, Max-sum, Max-min, LexRank, Biased LexRank, DivRank, and Grasshopper, aligns with my project's strategy. The rigorous evaluation framework and use of a common law dataset provide a robust methodology, guiding my own experimental design. The paper's insights, particularly the superiority of web search diversification techniques, offer valuable considerations for achieving a balance between relevance and diversity in personalized language models, shaping the foundation of my project's approach.

[5] The integration of the "Portfolio Theory of Information Retrieval" has significantly shaped and enhanced my project on personalized language models (LLMs) optimization. Drawing inspiration from the Modern Portfolio Theory, the paper's emphasis on strategically composing a portfolio of relevant documents, considering both mean relevance and variance as a measure of risk, has provided a nuanced perspective for document ranking under uncertainty. The introduced efficient ranking algorithm, accounting for uncertainties and correlations among retrieved documents, aligns seamlessly with the project's goal of balancing relevance and diversity in search outcomes. The paper's quantification of the benefits of diversification further resonates with the project's objective, promising to contribute to the advancement of personalized language models and optimized search experiences.

.

3. PROPOSED METHOD

Our methodology encompasses a comprehensive workflow that involves data extraction, preprocessing, and experimentation across various personalized tasks. The research question investigates the efficacy of applying Maximal Marginal Relevance (MMR) to enhance the relevance and diversity of BM25 results in the context of personalized language models.

3.1 DATASET

We curated three datasets—Personalized Citation Identification, Personalized News Categorization, and Personalized Scholarly Title Generation—to rigorously assess the integration of BM25 and MMR in personalized tasks. The first dataset focused on citation identification, aligning with BM25's information retrieval facet. Introducing MMR aimed to enhance both relevance and diversity in identified citations. The second dataset, Personalized News Categorization, sought to improve user-specific article categorization, highlighting MMR's potential to diversify news while maintaining relevance. The third dataset involved generating scholarly titles aligned with user interests, where MMR was anticipated to contribute to a diverse set of titles, enhancing the overall user experience.

Despite the project's overarching goal of investigating the effectiveness of diversifying BM25 results using MMR for personalized Language Models, not all datasets exhibited improvement. The absence of enhancement in the Personalized Citation Identification dataset could be attributed to the impact of diversification on initial BM25 results. The addition of diverse topics during diversification altered top-ranked results tied closely to the query, potentially affecting accuracy. Out of the six datasets (three user-based and three time-based), five showed little improvement, highlighting the nuanced impact of diversification across tasks. Surprisingly, the Personalized News Categorization dataset demonstrated significant improvement, emphasizing the task-specific nature of MMR's effectiveness in enhancing personalized information retrieval models.

3.2 Methodology Overview

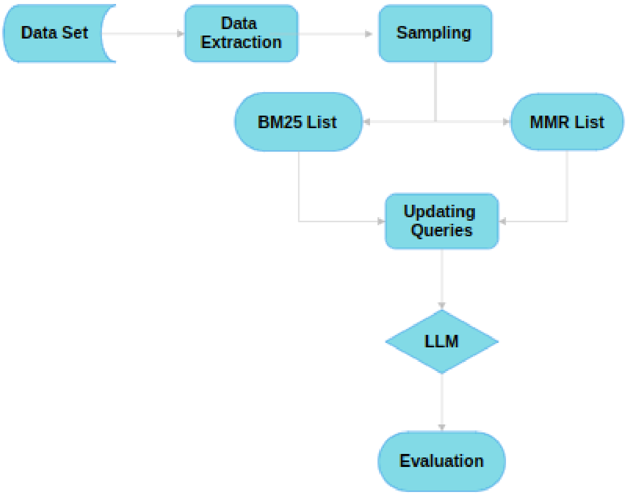
In the initial phase of our research, we concentrated on preprocessing of the validation dataset, aligning with our zero-shot approach that eliminated the necessity for model training. Leveraging GPT-3.5 Turbo as our large language model, we forewent any fine-tuning of parameters, streamlining our methodology for efficiency. Acknowledging resource constraints, we implemented a random data selection strategy in the subsequent step. From each of the three validation datasets, we randomly sampled 100 data points, ensuring a representative yet manageable subset for downstream analysis.

The application of the BM25 algorithm marked a pivotal point in our workflow. By running the BM25 code on the datasets, we obtained BM25 scores, and the top-k values from the resulting ranked list were seamlessly appended to the query. This strategic integration aimed to enhance query relevance and context. The refined queries were then fed into the large language model, GPT-3.5 Turbo, which generated outputs based on the personalized context embedded in the queries.

To ensure consistency and meet desired structural criteria, we employed the *output_formatting.py* script in the subsequent step. This script played a crucial role in organizing and structuring the generated outputs according to predefined standards. Further diversification of the BM25 results was achieved through the application of the MMR algorithm. Operating on the top 20 results, MMR altered the order of results, creating a more diverse and relevant subset for the subsequent query refinement. Additionally we have also experimented with three distinct lambda values [0.4,0.5,0.6] which are trade-off parameters that balances the relevance and diversity of the selected items.

$$MMR \overset{\text{def}}{=} Arg \max_{D_i \in R \setminus S} \left[\lambda (Sim_1(D_i, Q) - (1-\lambda) \max_{D_j \in S} Sim_2(D_i, D_j)) \right]$$

The conclusive step in our workflow involved the evaluation of the generated outputs. Accuracies between predicted outputs and actual outputs from the LaMP website were systematically assessed using the *eval_task.py* script. This evaluation step provided a quantitative measure of the performance of our methodology across different tasks and datasets, offering valuable insights into the effectiveness of our approach.



3.3 Experiment Workflow

In crafting this research paper, we designed an experiment workflow aimed at optimizing personalized language models without the need for model fine-tuning. Our systematic approach comprised six key steps, each contributing to the comprehensive evaluation of the language model's performance. The first step involved data extraction from the LaMP website, achieved through the implementation of a Python script named *data_set.py*. This script efficiently navigated relevant URLs on the LaMP platform, extracting datasets tailored to our experimental requirements. Recognizing resource constraints in Step 2, we employed the *random_100_datapoints.py* script to randomly select 100 data points from each dataset. This step ensured a manageable yet representative subset for subsequent analysis. To add personalization to the query in Step 3, we implemented the BM25 and MMR algorithms.

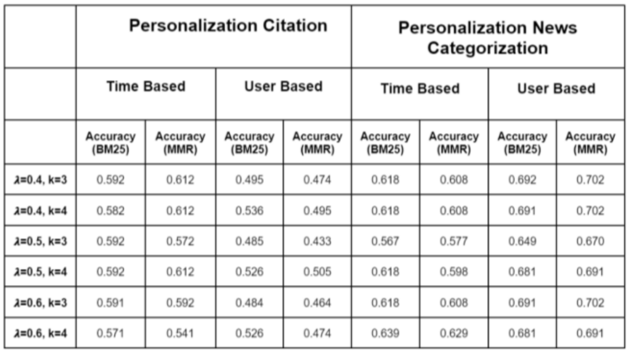
The BM25 results list was obtained, and the top-k results were integrated into the query. Subsequently, MMR was applied to diversify the top 20 BM25 results, creating a diversified subset for query personalization. In Step 4, the updated queries were fed into a large language model, specifically utilizing gpt-3.5-turbo, to generate outputs that captured the nuances of the personalized context. The generated outputs underwent formatting in Step 5, orchestrated by the *output_formatting.py* script. This crucial step ensured the outputs aligned with the desired structure, meeting predefined evaluation criteria.

The final step, Step 6, involved the evaluation of the formatted outputs against predefined criteria using the eval_task.py script. This rigorous evaluation process served as the cornerstone of our research, providing insights into the language model's performance across diverse tasks and datasets. Collectively, this experiment workflow not only addresses the challenges of personalized language models but also establishes a robust methodology for evaluating their effectiveness. The results and insights garnered from this research contribute to the broader conversation on optimizing language models for personalized contexts without resorting to extensive fine-tuning.

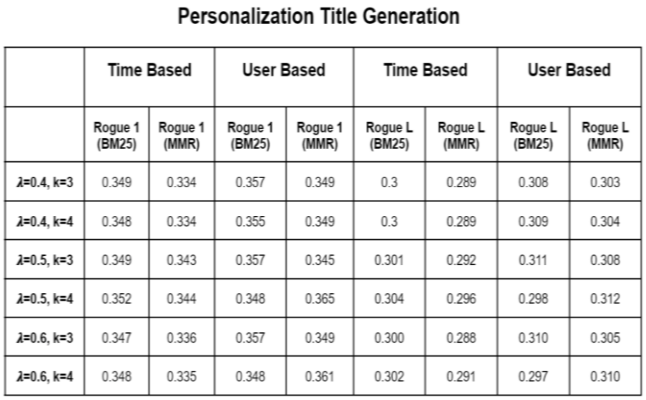
4. RESULTS AND DISCUSSIONS

In the validation of our research idea against six datasets—comprising three user-based and three time-based sets—we observed varying impacts of MMR diversification on personalized language models. Five datasets displayed marginal improvement, suggesting the nuanced nature of diversification's influence across different tasks. Unexpectedly, the Personalized News Categorization dataset(user based) exhibited significant enhancement, providing a notable exception to the overall trend. The absence of improvement in certain datasets prompts a closer examination of the intricacies involved. One potential explanation lies in the dynamic introduced by diversification on BM25 results, particularly evident in datasets where improvement was not observed. Diversification introduces changes to the initially top-ranked results, closely tied to the query topic. This alteration, induced by the infusion of new and diverse topics into the query, holds the potential to impact the final output, consequently affecting accuracy. This observation sheds light on the contextual sensitivity of diversification techniques and its varying impact on different datasets.

Further adding to the complexity of our findings, the experiments yielded mixed results across all three datasets. Interestingly, when assessing accuracy at different values of k (k=3 and k=4), the same conclusion over the K values has been observed in the LaMP paper[2] we noted that the accuracies were nearly identical, with only a few outliers deviating from this trend. This stability in accuracy across different values of k suggests a robust performance of the model, indicating that the choice of k does not significantly impact the overall effectiveness of our approach.Lastly, employing diversification for personalized large language models proves not ideal.

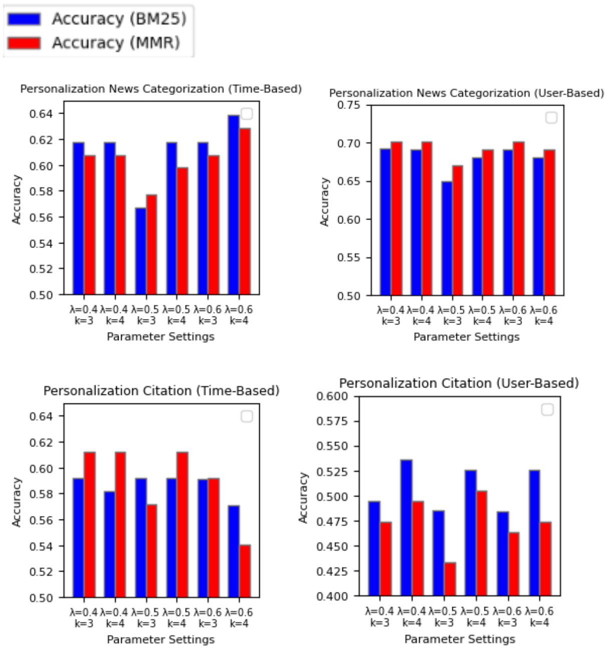


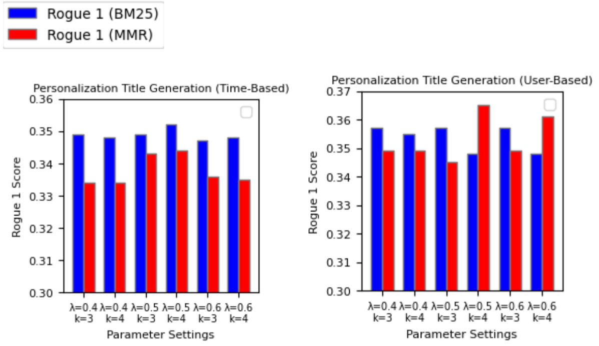
	Personalization Citation				Personalization News Categorization			
	Time Based		User Based		Time Based		User Based	
	Accuracy (BM25)	Accuracy (MMR)	Accuracy (BM25)	Accuracy (MMR)	Accuracy (BM25)	Accuracy (MMR)	Accuracy (BM25)	Accuracy (MMR)
λ=0.4, k=3	0.592	0.612	0.495	0.474	0.618	0.608	0.692	0.702
λ=0.4, k=4	0.582	0.612	0.536	0.495	0.618	0.608	0.691	0.702
λ=0.5, k=3	0.592	0.572	0.485	0.433	0.567	0.577	0.649	0.670
λ=0.5, k=4	0.592	0.612	0.526	0.505	0.618	0.598	0.681	0.691
λ=0.6, k=3	0.591	0.592	0.484	0.464	0.618	0.608	0.691	0.702
λ=0.6, k=4	0.571	0.541	0.526	0.474	0.639	0.629	0.681	0.691



Personalization Title Generation

	Time Based		User Based		Time Based		User Based	
	Rogue 1 (BM25)	Rogue 1 (MMR)	Rogue 1 (BM25)	Rogue 1 (MMR)	Rogue L (BM25)	Rogue L (MMR)	Rogue L (BM25)	Rogue L (MMR)
λ=0.4, k=3	0.349	0.334	0.357	0.349	0.3	0.289	0.308	0.303
λ=0.4, k=4	0.348	0.334	0.355	0.349	0.3	0.289	0.309	0.304
λ=0.5, k=3	0.349	0.343	0.357	0.345	0.301	0.292	0.311	0.308
λ=0.5, k=4	0.352	0.344	0.348	0.365	0.304	0.296	0.298	0.312
λ=0.6, k=3	0.347	0.336	0.357	0.349	0.300	0.288	0.310	0.305
λ=0.6, k=4	0.348	0.335	0.348	0.361	0.302	0.291	0.297	0.310





The introduction of diversification brings modifications to the initially highest-ranked results, closely related to the query topic. This change, resulting from the inclusion of new and varied topics in the query, has the capacity to influence the ultimate output, thereby impacting accuracy. This insight highlights the contextual responsiveness of diversification techniques and underscores its diverse effects on different datasets.

5. CONCLUSION

In investigating "MMR-Diversified BM25 for Personalized Language Models Optimization," our study uncovered nuanced findings. While BM25 demonstrated its expected effectiveness, the application of MMR diversification to personalized language models presented a mixed landscape. Contrary to conventional wisdom, MMR did not universally enhance results. The interaction between diversification and personalized context introduced complexities that impacted accuracy. Our project emphasizes the need for a nuanced, dataset-specific approach to optimizing personalized language models, revealing both challenges and potential improvements in the pursuit of contextually relevant query outputs.

6. REFERENCES

[1] J. Carbonell and J. Stewart, "The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries," SIGIR Forum, vol. 34, no. 1, 1999, Article 25, pp. 10.1145/290941.291025.

[2] A. Salemi, S. Mysore, M. Bendersky, and H. Zamani, "LaMP: When Large Language Models Meet Personalization," 2023.

[3] A. Kumar, "Maximal Marginal Relevance to Re-rank Results in Unsupervised KeyPhrase Extraction," Tech That Works, Oct. 24, 2019. [Online]. Available: https://medium.com/tech-that-works/maximal-marginal-relevance-to-rerank-results-in-unsupervised-keyphrase-extraction-22d95015c7c5

[4] M. Koniaris, I. Anagnostopoulos, and Y. Vassiliou, "Evaluation of Diversification Techniques for Legal Information Retrieval," Algorithms, vol. 10, no. 1, 2017, Art. no. 22, doi: 10.3390/a10010022.

[5] J. Wang and J. Zhu, "Portfolio Theory of Information Retrieval," in Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09), Association for Computing Machinery, New York, NY, USA, 2009, pp. 115–122.