Overview of the PIR Track at FIRE 2024: Evaluation of Personalised Information Retrieval

Pranav Kasela[1], Marco Braga[1,*], Effrosyni Sokli[1], Gian Carlo Milanese[1], Georgios Peikos[1], Sandip Modha[1], Alessandro Raganato[1], Marco Viviani[1] and Gabriella Pasi[1]

[1]*Information and Knowledge Representation, Retrieval, and Reasoning (IKR3) Lab, Department of Informatics, Systems, and Communication (DISCo) University of Milano-Bicocca, Milan, Italy*

Abstract

This abstract provides a short overview of the first edition of the shared task on Personalised Information Retrieval (PIR) organized at the 16th Forum for Information Retrieval Evaluation (FIRE 2024). A more detailed discussion of the approaches used by the participating teams is available in the track overview paper. PIR 2024 consisted of two sub-tasks. The first sub-task aims to explore the personalisation in cQA based on user profiles, following the standard IR pipeline. The second one, instead, aims to investigate the personalisation in cQA based on user profiles using recent LLMs and prompt engineering. Although the tasks saw an enthusiastic response in registrations, with 10 teams requesting the dataset, only 1 team finally submitted the runs, and 2 of them submitted the working notes.

Keywords

Information Retrieval, Question Answering, Personalization, Large Language Model

1. Introduction

The aim of the proposed shared tasks in FIRE 2024 is to elaborate on activities in the evaluation of Personalised Information Retrieval (PIR). Personalisation in Information Retrieval (IR) remains an important topic both in research and the development of practical applications for Information Retrieval. However, suitable means of evaluation of PIR remain an underexplored topic. PIR-FIRE aims to bring together researchers with a common interest in developing and evaluating novel methods for personalised operation of novel information retrieval applications.

Personalisation [1, 2, 3] and other adaptation of the search experience [4] to the user and search context is an important topic in IR, studied by the research community for a long time. Personalised search aims to tailor the search outcome to a specific user (or group of users) based on the knowledge of their interests and online behaviour. Unfortunately, text collections shared by initiatives on search evaluation do not usually provide the information needed for evaluating personalisation, i.e., information about specific users and their preferences. At the same time, the field has witnessed a transformation with the recent availability of pre-trained Large Language Models (LLMs). Despite recent research that has emphasized the advantages and drawbacks of personalizing LLMs [5], the development and evaluation of LLMs specifically designed to generate personalized responses remains inadequately explored.

For this reason, we propose running the PIR-FIRE laboratory shared task with the goal of organizing and introducing a new benchmark dataset coupled with user information allowing the development of personalised approaches, putting forward two sub-tasks: (i) Personalised Information Retrieval, (ii) Personalised Prompt-based Information Retrieval.

2. Task Definition

The first edition of PIR consisted of the following two sub-tasks:

Forum for Information Retrieval Evaluation (FIRE 2024), December 12–15, 2024, Gandhinagar, India

[*] Corresponding author.

m.braga@campus.unimib.it (M. Braga)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

2.1. Task 1: Standard IR

The cQA task will be tackled as a standard ad-hoc IR task, where the questions are going to be considered as the queries, and the collection, from which the answers will be retrieved, is composed by all the answers available in the dataset. In this case, personalization can be tackled using any standard or novel technique to create a user profile and inject it in the retrieval model. We plan to provide multiple baselines that utilize, as first stage retrievers, both classical approaches such as BM25, and neural approaches based on BERT-like models [6]. As a second stage, we plan to provide re-rankers, using cross-encoders, like Mono-T5 [7], for non-personalized baselines, and for personalized baselines, using of a mix of tags and historical documents related to the users and weighted according to their importance for the current question.

2.2. Task 2: Prompt-based IR

Differently from the second stage of the standard IR task, the proposed prompt-based baselines personalise the results by using models like Phi [8] and GPT [9] with prompts similar to the following one: "To which degree between 0 and 1 does the document [DOCUMENT] answer the question [QUESTION], and is relevant to a user with the following profile [USER PROFILE]", where the [USER PROFILE] is a series of user interests that are inferred from their activities and ordered according to their timestamp (most recent first) and importance.

More details about how this dataset was created can be found in the original resource paper [1].

3. Dataset and Evaluation

In this section, we discuss the datasets for each sub-task and the evaluation metrics used for each of them.

The PIR-FIRE will use data from StackExchange, a very popular community Question Answering (cQA) platform. The data is publicly available[1] under a cc-by-sa 4.0 license. The dataset is composed of questions, and their answers, collected from fifty communities, which can be categorized under the large umbrella of humanistic communities. In Table 1 we report the basic statistics for the dataset. Specifically: *document length*, measured in the number of words, *document score*, which is the difference between the number of up- and down-votes assigned by the community; *answers' count*, the number of answers given to a question; *comments' count*, the number of user comments to a given question or answer; *favorite count*, that indicates the number of users that flagged the question as their favorite, showing their interest in that topic; *tags count*, the number of tags associated to the question by the asking user.

The dump is curated and merged to tackle the cQA task as a retrieval task. The PIR-FIRE test collections provide the traditional components used in IR experiments, i.e. access to a document collection, search topics, and corresponding relevance judgments. Regarding the judgments, we only consider relevant the single answer that is explicitly labelled as the best answer by the user who submitted the question. In addition, our PIR evaluation test collections [1] are accompanied by user-related information for modelling and introducing profiles in evaluation experiments. The user-related information includes the text and the number of views of documents they have generated, and in many cases also the tags associated with these documents, the date since they are registered on the website, the badges they obtained, their reputation score, and some times also their autobiography. The information collected as previously explained can be used for personalising and adapting the search process to the current user, e.g. by creating and exploiting personal user profiles.

[1] https://zenodo.org/records/10679181

Table 1
Basic feature statistics for question and answers.

	Type	mean	std	median	25%	75%	99%
Questions	Length	125.69	112.90	94	60	153	553
	Score	5.13	10.73	2	1	6	45
#docs =	Answers Count	1.93	1.92	1	1	2	9
1,125,407	Comments Count	2.78	3.37	2	0	4	15
	Favorite Count	2.07	4.88	1	1	2	16
	Tags Count	2.45	1.21	2	1	3	5
Answers	Length	178.15	210.09	117	61	218	1000
#docs =	Score	5.13	12.43	2	1	5	51
2,173,139	Comments Count	1.62	2.62	1	0	2	12

Table 2
Results for both Tasks: we compare the proposed baselines with the runs submitted.

Model	P@1	NDCG@3	NDCG@10	R@100	MAP@100
BM25 (baseline)	0.279	0.353	0.394	0.707	0.362
BM25 + TAG (baseline)	0.306	0.383	0.425	0.707	0.392
BM25 + DistilBERT (baseline)	0.351	0.437	0.478	0.707	0.441
BM25 + + DistilBERT (Word Wizards)	0.315	0.396	0.439	0.707	0.403
BM25 + DistilBERT + TAG (baseline)	0.375	0.460	0.500	0.707	0.463
BM25 + DistilBERT + TAG (Word Wizards)	0.339	0.422	0.465	0.707	0.428
BM25 + MiniLM (baseline)	0.403	0.491	0.525	0.707	0.490
BM25 + MiniLM + TAG (baseline)	0.426	0.512	0.543	0.707	0.509
BM25 + T5-small (baseline)	0.376	0.469	0.506	0.707	0.468
BM25 + T5-small + TAG (baseline)	0.400	0.491	0.525	0.707	0.489
BM25 + T5-base (baseline)	0.417	0.517	0.548	0.707	0.510
BM25 + T5-base + TAG (baseline)	**0.440**	**0.535**	**0.563**	0.707	**0.528**

3.0.1. Evaluation setup

We will provide the participants with several baselines, including keyword and dense-based representations of user profiles (anonymised information gathered about individual users) as part of our data collection. For this shared task, we will use traditional evaluation metrics in the IR literature that can be applied also to personalized search, Precision (P), Recall (R), Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), and (normalized) Discounted Cumulative Gain (nDCG).

4. Participation

A total of 10 teams registered across both sub-tasks. Only one team (Word Wizards) submitted runs for task 1. For both tasks, 2 teams submitted the working notes.

Table 2 shows the performance of our proposed baselines and submitter runs.

5. Conclusion

The Personalised Information Retrieval (PIR) track at FIRE'24 focus on the evaluation of Personalised Information Retrieval (PIR), which remains an important topic both in research and the development of practical applications.

In future efforts, we plan to address the challenges encountered by making datasets more manageable, enhancing promotion, and broadening the scope of personalization to include diverse tasks in Information Retrieval (IR) and possibly Natural Language Processing (NLP). By providing resources such as pre-trained models, smaller datasets, and novel tasks, we aim to encourage a stronger focus on personalization across varied domains. These steps will help attract a broader range of participants and methodologies, driving greater engagement and impact.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] P. Kasela, M. Braga, G. Pasi, R. Perego, Se-pqa: Personalized community question answering, in: Companion Proceedings of the ACM on Web Conference 2024, WWW '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 1095–1098. URL: https://doi.org/10.1145/3589335.3651445. doi:10.1145/3589335.3651445.

[2] M. Braga, Personalized large language models through parameter efficient fine-tuning techniques, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024, pp. 3076–3076.

[3] M. Braga, P. Kasela, A. Raganato, G. Pasi, Synthetic data generation with large language models for personalized community question answering, arXiv preprint arXiv:2410.22182 (2024).

[4] P. Kasela, G. Pasi, R. Perego, N. Tonellotto, Desire-me: Domain-enhanced supervised information retrieval using mixture-of-experts, in: European Conference on Information Retrieval, Springer, 2024, pp. 111–125.

[5] A. Salemi, S. Mysore, M. Bendersky, H. Zamani, Lamp: When large language models meet personalization, 2023. arXiv:2304.11406.

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

[7] R. Nogueira, Z. Jiang, R. Pradeep, J. Lin, Document ranking with a pretrained sequence-to-sequence model, in: T. Cohn, Y. He, Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 708–718. URL: https://aclanthology.org/2020.findings-emnlp.63. doi:10.18653/v1/2020.findings-emnlp.63.

[8] M. Abdin, S. A. Jacobs, A. A. Awan, J. Aneja, A. Awadallah, H. Awadalla, N. Bach, A. Bahree, A. Bakhtiari, H. Behl, et al., Phi-3 technical report: A highly capable language model locally on your phone, arXiv preprint arXiv:2404.14219 (2024).

[9] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).