SE-PQA: StackExchange Personalized Community Question Answering*

Pranav Kasela[1,2], Marco Braga[1,3,*], Gabriella Pasi[1] and Raffele Perego[2]

[1]Università degli Studi di Milano-Bicocca, Milano
[2]ISTI-CNR, Pisa, Italy
[3]Politecnico di Torino, Dipartimento di Automatica e Informatica DAUIN, Corso Duca degli Abruzzi, Torino

Abstract

Personalization in Information Retrieval (IR) is a topic studied by the research community since a long time. Nevertheless, the availability of high-quality, real-world datasets for large-scale experiments and model evaluation remains limited. This paper helps to fill this gap by introducing SE-PQA (StackExchange - Personalized Question Answering), a new curated dataset designed for the development and evaluation of personalized models in the domain of community Question Answering (cQA). SE-PQA encompasses over one million queries and two million answers, annotated with a rich set of features that capture the social interactions among users on a cQA platform. We provide reproducible baseline methods for the cQA task based on the resource, including deep learning and personalized approaches. The results of the preliminary experiments conducted show the appropriateness of SE-PQA to train effective cQA models; they also show that personalization remarkably improves the effectiveness of all the methods tested.

Keywords

Question Answering, User Model, Personalization, Resource

1. Introduction

The problem of Personalization in Information Retrieval (IR) has been explored by the research community since a long time [2, 3, 4, 5, 6, 7]. Personalized search tries to tailor search results to individual users or groups based on their interests and online behaviour. One of the biggest issues in the training of Personalized neural models is the lack of large-scale, publicly available datasets that include detailed user-related information. Common datasets, like the AOL query log [8], the Yandex query log and the CIKM Cup 2016 dataset are frequently used, even if they come with privacy concerns and limitations due to anonymization. Our proposed research SE-PQA (StackExchange - Personalized Question Answering) is specifically designed to develop and assess personalized models in community Question Answering (cQA) task. SE-PQA is based on StackExchange, a cQA platform encompassing 178 open forums. The dataset, derived from a publicly available dump of user-contributed content under a cc-by-sa 4.0 licence, includes around one million questions and two million answers, annotated with features that reflect user interactions. These features include vote counts, view numbers, favourite selections, topic tags, and user comments. Additionally, we provide comprehensive user profiles linked to their historical questions and answers, social biographies, reputation scores, and view counts. In this study, we adapt the cQA task to an ad-hoc retrieval scenario, wherein the question is treated as a query, and the answers are retrieved from a pool of indexed past answers. In this setup, the objective is to retrieve a ranked set of documents containing the most appropriate answers to the user's query. The dataset is shared according to the conditions detailed in the included license agreement[1]. The code to create the dataset and reprodcue the baselines is publicly available [2].

IIR2024: 14th Italian Information Retrieval Workshop, 5th - 6th September 2024, Udine, Italy
*This is an extended abstract of [1].
*Corresponding author.
pranav.kasela@unimib.it (P. Kasela); m.braga@campus.unimib.it (M. Braga); gabriella.pasi@unimib.it (G. Pasi); raffaele.perego@isti.cnr.it (R. Perego)

[1]https://zenodo.org/records/10679181
[2]https://github.com/pkasela/SE-PQA

Table 1
Comparison between SE-PQA and other English text-based datasets for personalized IR.

Dataset	Documents	Train Queries	Val Queries	Test Queries	Users	Avg. user docs	Avg. # relevant	Rel. Assessment
AOLIA[10]	1 291 695	212 386	31 064	36 052	30 166	136.62 ± 134.17	1.15 ± 0.46	inferred from click
PERSON[11]	616 889	-	-	-	558 898	-	5.5 ± 5.3	inferred from citation
MAG Computer Science[12]	4 809 684	552 798	5 583	6 497	5 260 279	61.94 ± 60.32	3.25 ± 3.27	inferred from citations
MAG Physics	4 926 753	728 171	7 355	6 366	5 835 016	60.98 ± 56.54	4.17 ± 4.15	inferred from citations
MAG Political Science	4 814 084	162 597	1 642	5 715	6 347 092	40.64 ± 29.32	3.88 ± 5.17	inferred from citations
MAG Psychology	4 215 384	544 882	5 503	12 625	4 825 578	61.66 ± 62.72	4.73 ± 4.4	inferred from citations
Amazon Electorincs[13]	1 689 188	904	-	85	192 403	8.78 ± 8.26	1.12 ± 0.48	synthetic
Amazon Kindle Store	982 618	3 313	-	1 290	68 223	35.65 ± 37.48	1.87 ± 3.3	synthetic
Amazon CDs	1 097 591	534	-	160	75 258	21.75 ± 16.53	2.57 ± 6.59	synthetic
Amazon Cell Phones	194 439	134	-	31	27 879	4.95 ± 2.6	1.52 ± 1.13	synthetic
SE-PQA [Base]	2 073 370	822 974	78 854	99 878	588 688	34.71 ± 107.33	2.07 ± 1.81	inferred from answer scores
SE-PQA [Pers]	2 073 370	224 366	18 086	19 811	17 963	75.67 ± 142.41	1	manually selected by user

2. The Proposed Resource

Data from StackExchange has been utilized in the training of language models for sentence similarity [9]. However, to the best of our knowledge, the application of StackExchange for Question and Answer (Q&A) tasks has been limited to the selection of similar sentence training pairs, without using user or social features for personalized information retrieval tasks. With SE-PQA, we address previous limitations by offering a comprehensive, curated dataset comprising textual questions and answers from diverse, heterogeneous forums. In SE-PQA, a user can belong to multiple communities: for instance, by considering users who have contributed at least five documents (either questions or answers), we observe that out of the resulting 62,000 users, only 37% contributed to just a single community and, for example, 28% wrote documents in more than three communities. To enhance diversity, SE-PQA integrates data from multiple networks that fall under the broad category of humanistic communities. These communities, while focusing on different topics, utilize language that is not excessively varied among them. The dataset comprises a total of 1,125,407 questions, of which 1,001,706 (89%) have at least one answer. The training, validation, and test splits are performed temporally to prevent data leakage.

We infer the relevance of an answer to a question based on the number of up-votes it receives from community members. For experiments involving personalized cQA models, we consider only the single answer explicitly labelled as the best answer by the user who submitted the question as relevant. Using this information, we define two versions of the dataset: the *base* version, where we consider all answers with a positive score as relevant for a question, and the personalized (*pers*) version, which considers only the single answer selected by the user as the best answer as relevant for both the user and the question. A variety of user-generated information from the training set can be utilized in the personalization phase. For each question, we include all prior user posts to prevent data leakage during the training. However, user data is not limited to these documents: it also encompasses social interactions between users, tags assigned by users to previous questions along with their meanings, and badges earned by users. Additionally, the dataset contains biographic text (*about me*) introducing each user, a comprehensive set of numeric features (e.g., user reputation score, number of up-votes and down-votes for each post, number of views), and temporal information (e.g., user creation date, last access date, post creation timestamp).

Comparison with available datasets. In Table 1 we summarize the basic statistics of the main datasets used in the literature for personalized IR tasks. The AOL query log, released in 2006, remains a widely used resource despite privacy concerns, containing around 20 million web queries from over 657,000 users. However, the dataset's limitation lies in representing web pages solely by their URLs without text content. To address this, researchers have used a 2017 version that includes scraped text content, though this approach introduces the issue of web page content changes over time. The more recent AOLIA dataset resolves this by retrieving 2006 web page versions from the Internet Archive, providing a cleaner and higher-quality query set. Additionally, synthetic datasets such as PERSON, Amazon product search, and MAG have been developed to circumvent real-world dataset limitations. PERSON and MAG use citation networks to create personalized evaluation frameworks, while the

Amazon product search dataset employs item categories and properties to generate user queries. As depicted in Table 1, the proposed dataset exhibits similarities with other datasets concerning corpus volume and other statistical measures. Notably, it stands out as the largest dataset in terms of the number of queries provided. Unique to this dataset is the explicit annotation by users for relevance assessment. A single user labels the best answer, while various community members contribute by either up-voting or down-voting an answer based on its perceived relevance.

3. Experimental Evaluation

Table 2
Results for the cQA task on Base SE-PQA.

Model	P@1	NDCG@3	NDCG@10	MAP@100
BM25	0.330	0.325	0.359	0.320
BM25 + TAG	0.355*	0.349*	0.383*	0.342
BM25 + DistilBERT	0.404	0.400	0.435	0.389
BM25 + DistilBERT + TAG	0.422*	0.415*	0.448*	0.402*
BM25 + T5-small	0.448	0.442	0.471	0.426
BM25 + T5-small + TAG	0.463*	0.454*	0.482*	0.436*
BM25 + MiniLM	0.473	0.459	0.486	0.443
BM25 + MiniLM + TAG	**0.493***	**0.475***	**0.500***	**0.457***
BM25 + T5-base + TAG	**0.497**	**0.491**	**0.514**	**0.470**

Table 3
Results for the cQA task on Pers SE-PQA.

Model	P@1	NDCG@3	NDCG@10	MAP@100
BM25	0.279	0.353	0.394	0.362
BM25 + TAG	0.306*	0.383*	0.425*	0.392*
BM25 + DistilBERT	0.351	0.437	0.478	0.441
BM25 + DistilBERT + TAG	0.375*	0.460*	0.500*	0.463*
BM25 + T5	0.376	0.469	0.506	0.468
BM25 + T5 + TAG	0.400*	0.491*	0.525*	0.489*
BM25 + MiniLM	0.403	0.491	0.525	0.490
BM25 + MiniLM + TAG	**0.426***	**0.512***	**0.543***	**0.509***
BM25 + T5-base	0.417	0.517	0.548	0.510
BM25 + T5-base + TAG	**0.440***	**0.535***	**0.563***	**0.528***

In this section, we briefly describe the experimental setup and discuss the results of the preliminary experiments conducted. We adopt a two-stage ranking architecture: the first stage uses BM25 as a fast ranker; for the second stage, we rely on a linear combination of the scores computed by BM25, a neural re-ranker based on a pre-trained language model, and, when used, a personalization model exploiting user history, represented by the tags used by the users. In the second stage, three neural models are employed: MiniLM, which was trained and tuned using billions of training pairs, including StackExchange data; DistilBERT; MonoT5 small and base. For the DistilBERT and MonoT5 models, fine-tuning is performed using all the training queries of SE-PQA. To fine-tune MonoT5 model we rely on Adapter modules [14, 15, 16]. The intermediate dimension of the Adapter is set to 48. The personalization score is computed for an answer **a** generated in response to a query **q** written by user **u**, wherein the interests of **u** are captured through the set of tags assigned to all her/his previous questions posted before time **t**. The personalized score, called the TAG model, assesses the relevance of **a** to **q** by computing the intersection of tags associated with **a** and **u**'s previous questions, normalized by the total number of tags associated with **u**'s previous questions plus one, to account for cases where the set of tags is empty. We use P@1, NDCG@3, NDCG@10 and MAP@100 as our evaluation metrics. The experimental results are presented in Tables 2 and 3 for the *base* and *pers* datasets, respectively. Statistically significant improvements, indicated by *, are determined using a Bonferroni-corrected two-sided paired Student's t-test at a 99% confidence level. Neural re-rankers based on MiniLM outperform DistilBERT and T5-small, due to MiniLM's extensive training set. DistilBERT and T5-small, fine-tuned for 10 epochs, show MAP@100 improvements of approximately 22% and 33% over BM25, respectively. T5-base, on the other hand, outperforms all the baselines, obtaining relative improvements in terms of MAP@100 of 6% and 46% over MiniLM and BM25, respectively. The most notable result is that TAG improves, by a statistically significant margin, any cQA method it is combined with and for all the metrics considered on the *pers* version of the dataset, thus showing the advantages of personalization. The improvement due to the addition of this simple personalized model reaches up to 8% in terms of MAP@100 compared to their non-personalized baseline. We claim that personalization is particularly useful for multi-domain collections, where we can exploit information about users' interests in multiple topics of different domains. To validate our hypothesis, we perform a series of experiments considering single-domain data extracted from SE-PQA. Specifically, we consider 50 partitions of SE-PQA (base version) built by isolating the data from the 50 communities. We apply to each one of these subsets the non-personalized and personalized combinations of models using the bi-encoder model MiniLM, and measure the performance according to the same metrics used for the previous cQA tests. For

Table 4
Results for the cQA task on single-community data extracted from Base SE-PQA.

Community	Model (BM25 +)	P@1	NDCG@3	NDCG@10	R@100	MAP@100	λ
Academia	MiniLM	0.438	0.382	0.395	0.489	0.344	(.1,.9)
	MiniLM + TAG	**0.453***	**0.392***	**0.403***	0.489	**0.352***	(.1,.8,.1)
Apple	MiniLM	0.327	0.351	0.381	0.514	0.349	(.1,.9)
	MiniLM + TAG	**0.335***	**0.361***	**0.389***	0.514	**0.357***	(.1,.8,.1)
Bicycles	MiniLM	0.405	0.380	0.421	0.600	0.365	(.1,.9)
	MiniLM + TAG	**0.436***	**0.405***	**0.441***	0.600	**0.386***	(.1,.8,.1)
Christianity	MiniLM	0.534	0.505	0.555	0.783	0.497	(.2,.8)
	MiniLM + TAG	**0.549***	**0.521***	**0.564***	0.783	**0.507***	(.1,.8,.1)
Cooking	MiniLM	0.600	0.567	0.600	0.719	0.553	(.1,.9)
	MiniLM + TAG	**0.619***	**0.583***	**0.614***	0.719	**0.568***	(.1,.8,.1)
DIY	MiniLM	0.323	0.313	0.346	0.501	0.302	(.1,.9)
	MiniLM + TAG	**0.335***	**0.324***	**0.356***	0.501	**0.312***	(.1,.8,.1)
Hermeneutics	MiniLM	0.589	0.538	0.593	0.828	0.526	(.2,.8)
	MiniLM + TAG	**0.632***	**0.570***	**0.617***	0.828	**0.552***	(.1,.8,.1)
Law	MiniLM	0.663	0.647	0.678	0.803	0.639	(.2,.8)
	MiniLM + TAG	**0.677***	**0.657***	**0.687***	0.803	**0.649***	(.1,.8,.1)
Money	MiniLM	0.545	0.535	0.563	0.706	0.515	(.2,.8)
	MiniLM + TAG	**0.559***	**0.542***	**0.571***	0.706	**0.523***	(.1,.8,.1)
Music	MiniLM	0.508	0.447	0.476	0.602	0.418	(.2,.8)
	MiniLM + TAG	**0.522***	**0.460***	**0.486***	0.602	**0.427***	(.1,.8,.1)
Rpg	MiniLM	0.657	0.646	0.685	0.849	0.640	(.2,.8)
	MiniLM + TAG	**0.677***	**0.660***	**0.695***	0.849	**0.651***	(.1,.8,.1)
Scifi	MiniLM	0.532	0.563	0.596	0.745	0.559	(.2,.8)
	MiniLM + TAG	**0.549***	**0.574***	**0.606***	0.745	**0.569***	(.1,.8,.1)
$\lambda_{TAG} = 0$	english, health, history, travel, workplace, writers, woodworking, vegetarianism, skeptics, politics, philosophy, parenting, outdoors, musicfans, literature, linguistics, judaism, interpersonal, hsm, genealogy, freelancing, fitness, expatriates, buddhism, anime						
No Statistical Improvement	boardgames, gardening, gaming, hinduism, islam, lifehacks, martialarts, movies, opensource, pets, sound, sports, sustainability						

a fair comparison, we performed for each community the optimization of the λ weights on single-domain validation data. We notice that the contribution of the TAG model is lower in this setting, and in some cases missing. Specifically, for 25 out of 50 communities, personalization does not lead to any improvement, i.e., $\lambda_{TAG} = 0$. On the other 13 communities, we do not observe statistically significant improvements for P@1 over the non-personalized methods. As expected, the absolute metrics are slightly higher for single-domain tests due to the higher recall in the first-stage retrieval, since we drastically reduce the size of the collection indexed, allowing the first-stage ranker to perform better. However, in terms of the absolute performance boost due to the TAG model, we achieve a 2% improvement on P@1 when using all communities together, while the average boost decreases to 1.1% when considering the communities separately. In Table 4, we report the results for the 12 communities for which personalization achieves statistically significant improvements.

4. Conclusion and Future Work

Despite significant efforts by the IR community in studying personalization, a comprehensive dataset for evaluating and comparing different approaches has been lacking. This work addresses this gap by introducing a large-scale dataset encompassing 14 years of StackExchange user activity. Detailed information about the dataset is provided, along with its potential for training and evaluating both classical and personalized models for the community question-answering (cQA) task. Preliminary experiments demonstrate that personalization significantly enhances state-of-the-art methods based on pre-trained large language models. The analysis and unique features of the SE-PQA dataset suggest numerous future research directions, including the development of more complex personalized models utilizing additional user features not employed in the current models.

Acknowledgments

We acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources and support.

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] A. Borisov, I. Markov, M. de Rijke, P. Serdyukov, A context-aware time model for web search, in: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 205–214. URL: https://doi.org/10.1145/2911451.2911504. doi:10.1145/2911451.2911504.

[3] S. Calegari, G. Pasi, Personal ontologies: Generation of user profiles based on the yago ontology, Information Processing & Management 49 (2013) 640–658. URL: https://www.sciencedirect.com/science/article/pii/S0306457312001070. doi:https://doi.org/10.1016/j.ipm.2012.07.010, personalization and Recommendation in Information Access.

[4] M. Braga, A. Raganato, G. Pasi, et al., Personalization in bert with adapter modules and topic modelling, in: Proceedings of the 13th Italian Information Retrieval Workshop (IIR 2023). Pisa, Italy, 2023, pp. 24–29.

[5] E. Bassani, P. Kasela, G. Pasi, Denoising attention for query-aware user modeling, in: K. Duh, H. Gomez, S. Bethard (Eds.), Findings of the Association for Computational Linguistics: NAACL 2024, Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 2368–2380. URL: https://aclanthology.org/2024.findings-naacl.153. doi:10.18653/v1/2024.findings-naacl.153.

[6] 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, SIGIR '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 3076. URL: https://doi.org/10.1145/3626772.3657657. doi:10.1145/3626772.3657657.

[7] P. Kasela, G. Pasi, R. Perego, Se-pef: a resource for personalized expert finding, in: Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, SIGIR-AP '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 288–309. URL: https://doi.org/10.1145/3624918.3625335. doi:10.1145/3624918.3625335.

[8] G. Pass, A. Chowdhury, C. Torgeson, A picture of search, in: Proceedings of the 1st International Conference on Scalable Information Systems, InfoScale '06, Association for Computing Machinery, New York, NY, USA, 2006, p. 1–es. URL: https://doi-org.unimib.idm.oclc.org/10.1145/1146847.1146848. doi:10.1145/1146847.1146848.

[9] HuggingFace, Train a sentence embedding model with 1b training pairs, 2021. URL: https://huggingface.co/blog/1b-sentence-embeddings.

[10] S. MacAvaney, C. Macdonald, I. Ounis, Reproducing personalised session search over the aol query log, in: Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I, Springer-Verlag, Berlin, Heidelberg, 2022, p. 627–640. URL: https://doi.org/10.1007/978-3-030-99736-6_42. doi:10.1007/978-3-030-99736-6_42.

[11] S. A. Tabrizi, A. Shakery, H. Zamani, M. A. Tavallaei, Person: Personalized information retrieval evaluation based on citation networks, Information Processing & Management 54 (2018) 630–656. URL: https://www.sciencedirect.com/science/article/pii/S0306457317307811. doi:https://doi.org/10.1016/j.ipm.2018.04.004.

[12] E. Bassani, P. Kasela, A. Raganato, G. Pasi, A multi-domain benchmark for personalized search evaluation, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, CIKM '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 3822–3827. URL: https://doi.org/10.1145/3511808.3557536. doi:10.1145/3511808.3557536.

[13] Q. Ai, Y. Zhang, K. Bi, X. Chen, W. B. Croft, Learning a hierarchical embedding model for personalized product search, in: Proceedings of the 40th International ACM SIGIR Conference

on Research and Development in Information Retrieval, SIGIR '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 645–654. URL: https://doi-org.unimib.idm.oclc.org/10.1145/3077136.3080813. doi:10.1145/3077136.3080813.

[14] C. Poth, H. Sterz, I. Paul, S. Purkayastha, L. Engländer, T. Imhof, I. Vulić, S. Ruder, I. Gurevych, J. Pfeiffer, Adapters: A unified library for parameter-efficient and modular transfer learning, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Singapore, 2023, pp. 149–160. URL: https://aclanthology.org/2023.emnlp-demo.13.

[15] J. Pfeiffer, A. Kamath, A. Rücklé, K. Cho, I. Gurevych, AdapterFusion: Non-destructive task composition for transfer learning, in: P. Merlo, J. Tiedemann, R. Tsarfaty (Eds.), Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 487–503. URL: https://aclanthology.org/2021.eacl-main.39. doi:10.18653/v1/2021.eacl-main.39.

[16] M. Braga, A. Raganato, G. Pasi, AdaKron: An adapter-based parameter efficient model tuning with kronecker product, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 350–357. URL: https://aclanthology.org/2024.lrec-main.32.