PARK: Personalized Academic Retrieval with Knowledge-graphs*

Pranav Kasela[1,2,*], Gabriella Pasi[1] and Raffaele Perego[2]

[1]University of Milano-Bicocca, Italy
[2]ISTI-CNR, Pisa, Italy

Abstract

Academic Search is a search task aimed to manage and retrieve scientific documents like journal articles and conference papers. Personalization in this context meets individual researchers' needs by leveraging, through user profiles, the user related information (e.g. documents authored by a researcher), to improve search effectiveness and to reduce the information overload. While citation graphs are a valuable means to support the outcome of recommender systems, their use in personalized academic search (with, e.g. nodes as papers and edges as citations) is still under-explored.

Existing personalized models for academic search often struggle to fully capture users' academic interests. To address this, we propose a two-step approach: first, training a neural language model for retrieval, then converting the academic graph into a knowledge graph and embedding it into a shared semantic space with the language model using translational embedding techniques. This allows user models to capture both explicit relationships and hidden structures in citation graphs and paper content. We evaluate our approach in four academic search domains, outperforming traditional graph-based and personalized models in three out of four, with up to a 10% improvement in MAP@100 over the second-best model. This highlights the potential of knowledge graph-based user models to enhance retrieval effectiveness.

Keywords

Personalized information retrieval, Knowledge graphs, Neural information retrieval, Dense retrieval

1. Introduction

Academic search aims to retrieve relevant scientific documents, such as journal articles and conference papers, from large repositories, based on queries, generally, formulated by researchers, students or professionals with specific information need. While traditional retrieval systems focus on global relevance signals, they often overlook individual users' research profiles. Personalized academic search addresses this gap by tailoring results to each user's expertise and preferences. This task becomes particularly significant in domains like scientific research, where users develop long-term, domain-specific interests.

Despite increasing interest in personalized IR [2, 3, 4, 5, 6, 7, 8, 9, 10, 11], existing academic search methods underutilize structured bibliographic information such as citation networks to personalize search results. We propose **PARK** [1], *Personalized Academic Retrieval with Knowledge Graphs*, a novel framework for personalized academic search. PARK leverages citation graphs and neural language models to construct user embeddings and re-rank documents based on both semantic and structural similarity. PARK aligns the neural language models with knowledge graphs using a two-stage modeling strategy: (1) training a bi-encoder neural retrieval model for text relevance on an academic search dataset, and (2) learning knowledge graph embeddings from a citation-derived graph using TransE [12] and TransH [13]. By integrating these components, PARK achieves state-of-the-art performance in personalized retrieval across multiple academic domains.

IIR2025: 15th Italian Information Retrieval Workshop, 3th - 5th September 2025, Cagliari, Italy

*This is an extended abstract of [1].

*Corresponding author.
pranav.kasela@unimib.it (P. Kasela); gabriella.pasi@unimib.it (G. Pasi); raffaele.perego@isti.cnr.it (R. Perego)
https://pkasela.github.io/ (P. Kasela); http://raffaele.isti.cnr.it/ (R. Perego)
0000-0003-0972-2424 (P. Kasela); 0000-0002-6080-8170 (G. Pasi); 0000-0001-7189-4724 (R. Perego)

2. PARK Architecture

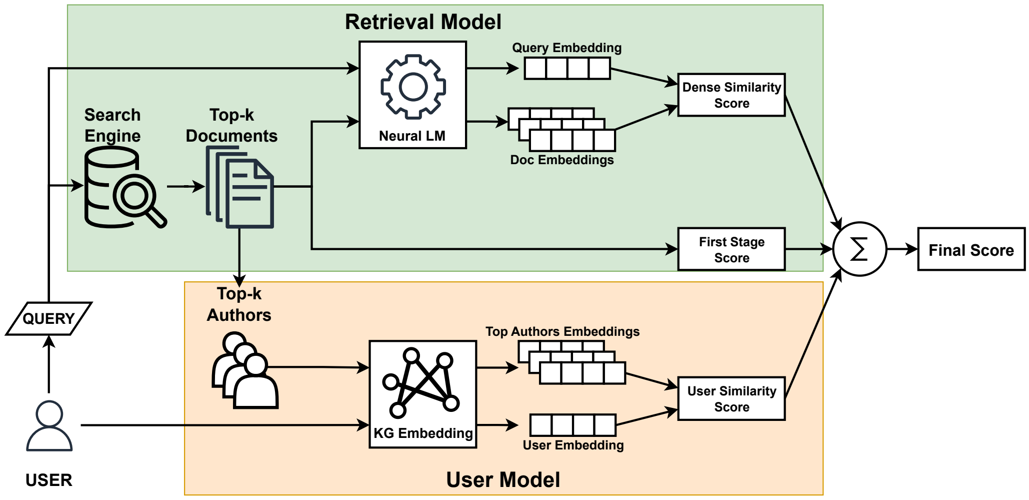

Figure 1: Overview of the PARK retrieval pipeline (Figure 1 from [1]).

As depicted in Figure 1, the architecture includes two main components:

- **Retrieval Model**: A two-stage IR pipeline using BM25 [14] for an efficient first stage document retrieval and a following second stage based on bi-encoder dense neural model (MiniLM) [15] for re-ranking.
- **User Model**: The proposed user model based on knowledge graph embeddings, which computes user similarity scores, reflecting the alignment of research profiles and interests of the user issuing the query and the authors of the documents retrieved by the first stage retriever.

The neural bi-encoder model is trained by minimizing the distance between the query representation and the associated relevant document representations while increasing the distance between the query representation and the non relevant documents representations using the Triplet Margin Loss [16]. The final score for each document is given by a weighted combination of:

- $\mathrm{BM25}(q, d)$: lexical similarity.
- $\mathrm{Dense}(q, d)$: semantic similarity from the neural model.
- $\mathrm{UserSim}(u, a_d)$: cosine similarity between the user embedding and the authors of document d in the knowledge graph embedding.

$$S(q, d) = \lambda_1 \cdot \mathrm{BM25}(q, d) + \lambda_2 \cdot \mathrm{Dense}(q, d) + \lambda_3 \cdot \mathrm{UserSim}(u, a_d) \tag{1}$$

with $\lambda_1 + \lambda_2 + \lambda_3 = 1$ and λ_i are optimized on a validation split.

3. User Modeling with Knowledge Graphs

PARK encodes the authors in a vector space shared with the language model in order to capture both bibliographic structure and textual semantics.

Academic Knowledge Graph Construction Starting from a citation graph (papers as nodes; citations as directed edges), we build a heterogeneous Knowledge Graph (KG) with four node types and five relations:

- **Nodes:** *Author* (user), *Document* (paper), *Venue* (conference or journal), and *Affiliation* (institution)
- **Relations:** wrote: Author → Document, cited: Document → Document, co_author: Author ↔ Author, in_venue: Document → Venue, and affiliated: Author → Affiliation

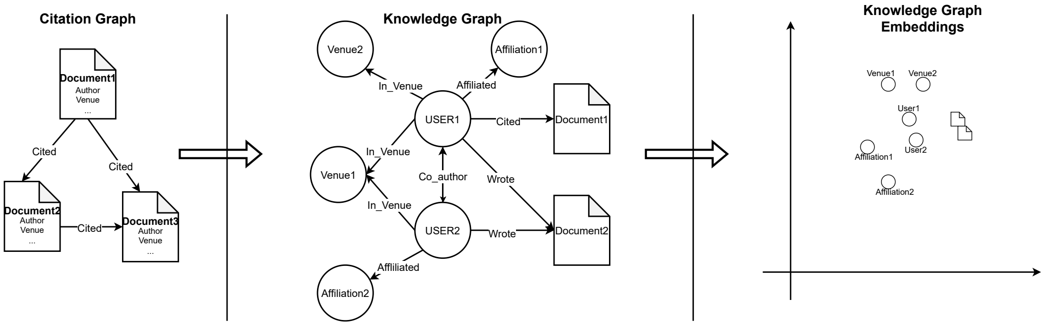

Figure 2: Overview of the process to obtain user embeddings from the citation Graph. The first step converts the Citation Graph to an Academic Knowledge Graph, and the second step embeds the Academic Knowledge Graph according to the proposed techniques.

Embedding Strategy We embed the KG into the same d-dimensional latent space as our neural retriever (MiniLM):

1. *Document nodes* are initialized with fixed embeddings from the pre-trained MiniLM encoder.
2. *Other nodes* (authors, venues, affiliations) and all relations are jointly embedded using: **TransE** [12] for PARK-E, **TransH** [13] for PARK-H

Fixing document embeddings preserves their semantic features, while TransE/TransH learn to position authors and entities relative to these fixed points, aligning structural properties of KG with the textual signals from the retrieval model.

User Embeddings & Scoring Each author u is represented by their learned KG embedding. At query time, we compute the user similarity score as the cosine similarity between the user embeddings of the query issuer and the authors of the documents being scored. The score reflects the alignment of the research profiles and interests of the user issuing the query and the authors of the paper being scored. This score is integrated into our final ranking formula (Eq. 1).

4. Evaluation

The system is evaluated on four datasets specifically designed for evaluating model in the context of personalized academic search (Computer Science, Political Science, Psychology, and Physics) [17]. We compare against the following baselines: BM25 [14], MiniLM [15], Mean [17], Attention [18], Self Citation, CrossEnc$_{\text{RA}}$ [3], CTRL$_{\text{It}}$ [19], PageRank [20], POP (Popularity) [17]. We evaluate with MAP@100, MRR@10, and NDCG@10. A convex sum of normalized scores ensures fair comparison across models. The code is made publicly available[1].

PARK outperforms all baselines in Political Science, Psychology, and Physics (Table 1). In Computer Science, the POP baseline remains competitive due to the high predictive value of citation counts. Overall, PARK demonstrates robust effectiveness across diverse disciplines.

5. Ablation Study

To understand the contribution of node types in the knowledge graph, we conduct an ablation study using PARK-H (Table 2). The goal was to evaluate the effect of each node type (user, venue, and affiliation) and their associated relations on the model's retrieval performance. Results in Table 2 indicate:

- Using only user nodes yields substantial gains over baseline.

[1]https://github.com/pkasela/PARK-Personalized_Academic_Retrieval_with_Knowledge-graphs

Table 1
Effectivess of PARK-E and PARK-H compared to the competing methods on the four datasets. The best-performing model is highlighted in boldface. Symbol * indicates a statistically significant difference over the second-best-performing model.

	Computer Science			Political Science			Psychology			Physics		
Model	MAP@100	MRR@10	NDCG@10	MAP@100	MRR@10	NDCG@10	MAP@100	MRR@10	NDCG@10	MAP@100	MRR@10	NDCG@10
BM25	0.123	0.489	0.225	0.133	0.502	0.241	0.126	0.512	0.239	0.128	0.537	0.269
MiniLM	0.193	0.600	0.301	0.186	0.580	0.297	0.218	0.647	0.342	0.183	0.624	0.335
Mean	0.199	0.606	0.308	0.193	0.598	0.306	0.220	0.652	0.347	0.189	0.639	0.345
Attention	0.201	0.612	0.312	0.199	0.612	0.314	0.220	0.656	0.349	0.190	0.648	0.348
Self Citation	0.213	0.624	0.325	0.205	0.613	0.321	0.237	0.689	0.370	0.204	0.671	0.365
CTRL_{lt}	-	0.629	0.322	-	0.648	0.338	-	0.685	0.370	-	0.667	0.366
$\text{CrossEnc}_{\text{RA}}$	-	0.635	0.324	-	0.651	0.338	-	0.700	0.380	-	0.673	0.369
PageRank	0.213	0.644	0.331	0.203	0.622	0.324	0.230	0.670	0.360	0.189	0.636	0.346
POP	**0.238***	**0.684***	**0.370***	0.214	0.649	0.345	0.225	0.656	0.356	0.206	0.670	0.370
PARK-E	0.228	0.651	0.344	0.232	0.661	0.356	**0.261***	**0.716***	**0.397***	**0.225**	0.695	**0.391**
PARK-H	0.230	0.655	0.346	**0.233***	**0.662***	**0.357***	0.255	0.712	0.392	**0.225***	**0.696***	**0.391***

- Adding venue nodes provides minimal additional improvement.
- Including affiliation nodes significantly boosts performance in all domains, confirming the value of institutional context.

Table 2
Ablation study results for each node type on four datasets.

	Computer Science			**Political Science**			**Psychology**			**Physics**		
Node Types	**MAP@100**	**MRR@10**	**NDCG@10**	**MAP@100**	**MRR@10**	**NDCG@10**	**MAP@100**	**MRR@10**	**NDCG@10**	**MAP@100**	**MRR@10**	**NDCG@10**
Only User	0.223	0.644	0.336	0.220	0.637	0.343	0.244	0.693	0.379	0.221	0.686	0.383
↳ + Venue	0.225	0.642	0.337	0.225	0.646	0.348	0.247	0.694	0.381	0.222	0.686	0.383
↳ + Affiliation	0.230	0.655	0.346	0.233	0.662	0.357	0.255	0.712	0.392	0.225	0.696	0.391

6. Discussion and Future Work

PARK demonstrates how knowledge graph-based user embeddings, when aligned with neural document encoders, can improve personalized academic search. By representing academic entities in a unified latent space, PARK captures both explicit citation relationships and latent author-topic patterns.

While PARK performs well across most domains, limitations remain. The use of fixed document embeddings may limit adaptability, and citation coverage biases could affect robustness. Future work will explore: softening constraints on document embeddings; integrating popularity-based priors for domains like CS; and adapting to open-world or streaming academic corpora.

Overall, PARK advances the state-of-the-art in personalized academic search by combining structured knowledge with dense text representations.

Acknowledgments

This work was supported by the European Union – Next Generation EU within the project NRPP M4C2, Investment 1.,3 DD. 341–15 march 2022 – FAIR – Future Artificial Intelligence Research – Spoke 4 - PE00000013 - D53C22002380006.
This work was partially supported by the Spoke “Human-centered AI” of the M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - “FAIR - Future Artificial Intelligence Research”, the “Extreme Food Risk Analytics” (EFRA) project, Grant no. 101093026, funded by European Union – NextGenerationEU.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4 in order to: Grammar and spelling check, Paraphrase and reword. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication’s content.

References

[1] P. Kasela, G. Pasi, R. Perego, Park: Personalized academic retrieval with knowledge-graphs, Information Systems (2025) 102574. URL: https://www.sciencedirect.com/science/article/pii/S0306437925000584. doi:https://doi.org/10.1016/j.is.2025.102574.

[2] 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.

[3] A. Salemi, S. Mysore, M. Bendersky, H. Zamani, LaMP: When large language models meet personalization, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 7370–7392. URL: https://aclanthology.org/2024.acl-long.399. doi:10.18653/v1/2024.acl-long.399.

[4] B. Tan, X. Shen, C. Zhai, Mining long-term search history to improve search accuracy, in: T. Eliassi-Rad, L. H. Ungar, M. Craven, D. Gunopulos (Eds.), Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006, ACM, 2006, pp. 718–723. URL: https://doi.org/10.1145/1150402.1150493. doi:10.1145/1150402.1150493.

[5] 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.

[6] K. Bi, P. Metrikov, C. Li, B. Byun, Leveraging user behavior history for personalized email search, in: Proceedings of The Web Conference 2021, 2021, pp. 3526–3537. URL: https://doi.org/10.1145/3442381.3450110.

[7] Q. Ai, D. N. Hill, S. V. N. Vishwanathan, W. B. Croft, A zero attention model for personalized product search, in: W. Zhu, D. Tao, X. Cheng, P. Cui, E. A. Rundensteiner, D. Carmel, Q. He, J. X. Yu (Eds.), Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019, ACM, 2019, pp. 379–388. URL: https://doi.org/10.1145/3357384.3357980. doi:10.1145/3357384.3357980.

[8] P. Kasela, M. Braga, G. Pasi, R. Perego, Se-pqa: Personalized community question answering, in: Companion Proceedings of the ACM 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.

[9] M. Braga, P. Kasela, A. Raganato, G. Pasi, Synthetic data generation with large language models for personalized community question answering, in: 2024 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2024, pp. 360–366. doi:10.1109/WI-IAT62293.2024.00057.

[10] S. Lu, Z. Dou, C. Xiong, X. Wang, J.-R. Wen, Knowledge enhanced personalized search, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 709–718. URL: https://doi.org/10.1145/3397271.3401089. doi:10.1145/3397271.3401089.

[11] 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.

[12] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, in: C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Weinberger (Eds.), Advances in Neural Information Processing Systems, volume 26, Curran Associates, Inc., 2013, p. 1.

[13] Z. Wang, J. Zhang, J. Feng, Z. Chen, Knowledge graph embedding by translating on hyperplanes,

Proceedings of the AAAI Conference on Artificial Intelligence 28 (2014). doi:10.1609/aaai.v28i1.8870.

[14] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, M. Gatford, Okapi at TREC-3, in: D. K. Harman (Ed.), Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994, volume 500-225 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 1994, pp. 109–126. URL: http://trec.nist.gov/pubs/trec3/papers/city.ps.gz.

[15] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, M. Zhou, Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 5776–5788. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[16] V. Balntas, E. Riba, D. Ponsa, K. Mikolajczyk, Learning local feature descriptors with triplets and shallow convolutional neural networks., in: Bmvc, volume 1, 2016, p. 3.

[17] 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. doi:10.1145/3511808.3557536.

[18] J. Jiang, T. Wu, G. Roumpos, H. Cheng, X. Yi, E. Chi, H. Ganapathy, N. Jindal, P. Cao, W. Wang, End-to-end deep attentive personalized item retrieval for online content-sharing platforms, in: WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020, ACM / IW3C2, 2020, pp. 2870–2877.

[19] S. Mysore, G. Dhanania, K. Patil, S. Kallumadi, A. McCallum, H. Zamani, Memory augmented cross-encoders for controllable personalized search, 2024. URL: https://arxiv.org/abs/2411.02790. arXiv:2411.02790.

[20] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web., Technical Report 1999-66, Stanford InfoLab, 1999. URL: http://ilpubs.stanford.edu:8090/422/, previous number = SIDL-WP-1999-0120.