A Multi-Domain Benchmark for Personalized Search Evaluation

Elias Bassani[1,2], Pranav Kasela[2], Alessandro Raganato[2], and Gabriella Pasi[2]

[1]Consorzio per il Trasferimento Tecnologico - C2T [2]University of Milano-Bicocca, Milan, Italy

Personalized Search Evaluation Issues

- **Datasets** shared by initiatives on search evaluation do **not** usually **provide** information about specific **users** and their **preferences**, which are needed for evaluating personalization.
- Some efforts have been devoted to defining large-scale task-related datasets for personalization: AOL Query Log, CIKM Cup 2016, and Yandex Query Log.
  - **Issues:** content availability, privacy concern, anonymized texts (*semantic retrieval approaches not usable*).
- Due to the above issues, researchers have proposed several methodologies to define synthetic datasets for Personalized Search Evaluation, most notably folksonomy-based datasets and Amazon reviews-based datasets.
  - **Issues:** low data and queries quality, low query availability, missing data, lack of validation.
- Another attempt to define a synthetic dataset for Personalized Search Evaluation is PERSON, a methodology based on citation networks proposed in "PERSON: Personalized information retrieval evaluation based on citation networks." by Tabrizi et al. (2018)
  - **Pros:** rich and large scale datasets, in-depth validation process;
  - **Cons:** lack of potentially valuable information for personalization from the original data, only one domain.
- There is still a **lack** of high-quality **benchmark** datasets for **Personalized Search evaluation**.

What We Propose

- We **revisit** and **extend** the **PERSON** methodology, overcoming the aforementioned limitations while keeping its benefits (*more details in the paper*).
- We **share** a large-scale **benchmark** across **four** academic **domains**, with more than **18 million documents** and **1.9 million queries**, designed for evaluating **Personalized Results Re-Ranking** approaches.
- We provide a **rich set of metadata** for each document, the data to derive the **user-document interactions**, pre-computed **BM25's result lists to re-rank**, and pre-computed **baseline runs**.
- Relations among the data, such as authorship relations and the paper references, can be represented by **graph structures** allowing graph-based personalization approaches.

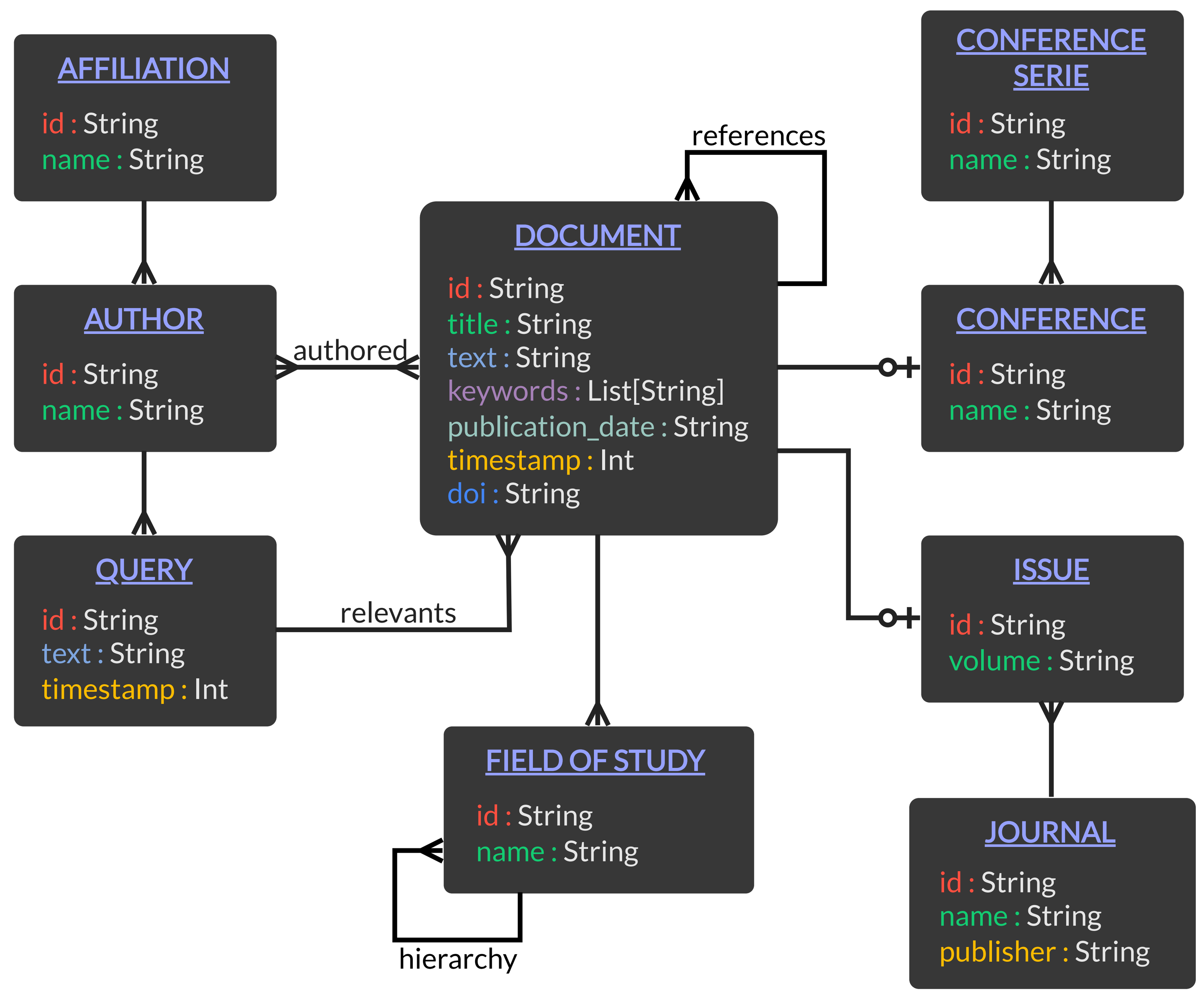
Opportunities

- The shared large-scale datasets are suited for **training** and **evaluating**:
  - **Content-based** personalization models.
  - **Collaborative filtering** approaches.
  - **Graph-based** personalization models.
  - **Joint Personalized Search** and **Recommendation** models.

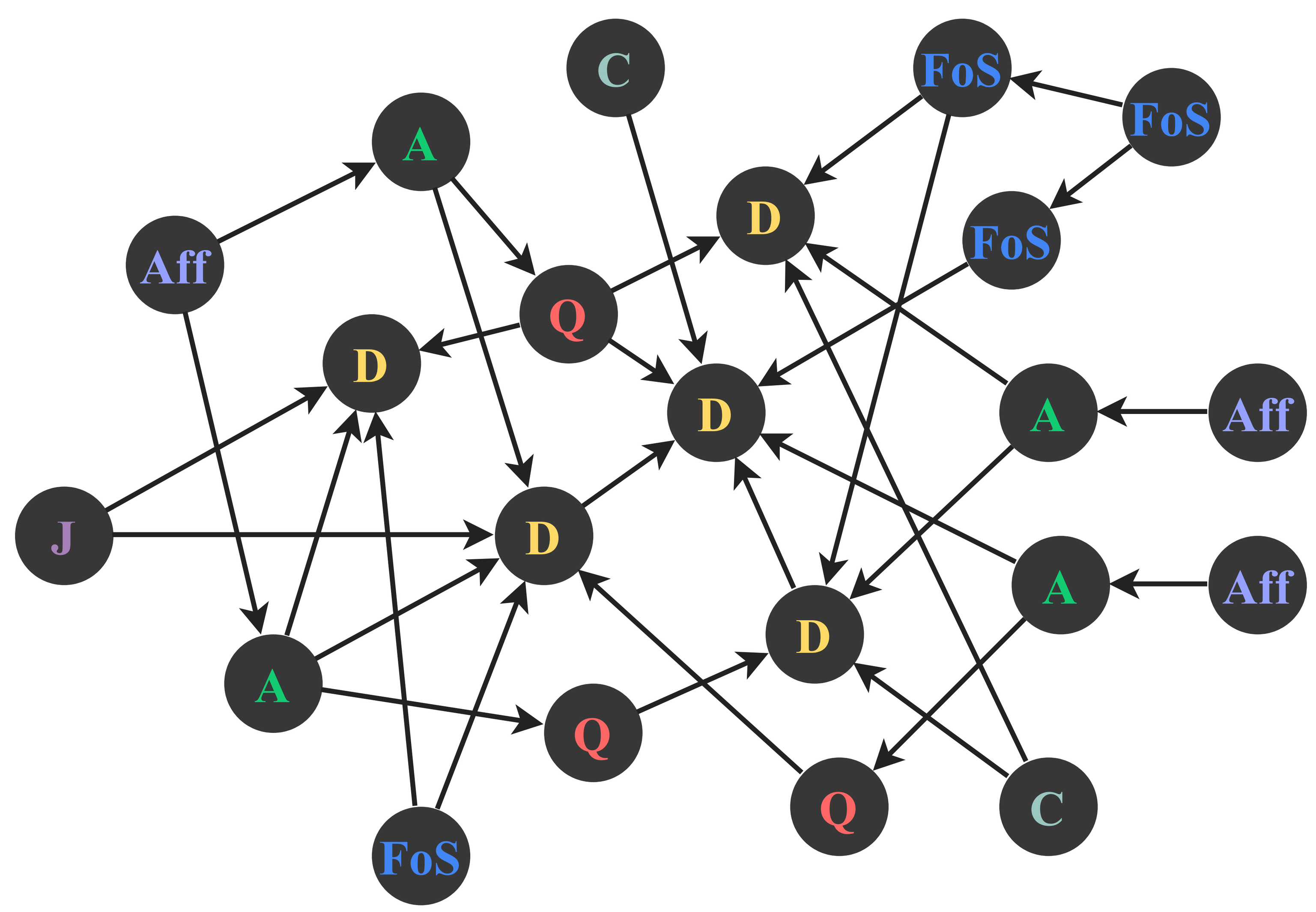
Query Generation

- **PERSON** methodology:
  - Starting point: academic papers.
  - title → query; author → user; references → relevant documents.
- Our **refinements**:
  - **Processing** to resemble real-world queries:
    - Stop-words removal; Non-destructive stemming (Krovetz stemmer)
  - Query **selection**:
    - Users must have at least 20 documents for personalization.
    - For personalization purposes, we consider only the papers published by the user/author before the one used as the query, thus preserving the temporal aspect.
    - Not all the references are necessarily relevant to the topic expressed by a paper's title. To reduce the presence of spurious relevant documents and malformed queries, we applied some heuristics based on BM25 results.

Structure of the Datasets



Graph Representation of the Datasets



Statistics of the Datasets

	Computer Science	Physics	Political Science	Psychology
# documents	4 809 684	4 926 753	4 814 084	4 215 384
# users	5 260 279	5 835 016	6 347 092	4 825 578
# train queries	552 798	728 171	162 597	544 882
# val queries	5 583	7 355	1 642	5 503
# test queries	6 497	6 366	5 715	12 625

Provided Baselines

- First-stage retriever:
  - **BM25:** Classic probabilistic retrieval model.
- Re-Rankers:
  - **Pop:** Popularity-based model.
  - **BiEnc:** Bi-encoder-based retrieval model.
  - **All-BiEnc:** BiEnc trained on all four domains to assess domain adaptation and transfer learning opportunities.
  - **Mean:** Embedding-based personalized retrieval model. It defines the user models as the average of their associated document representations.
  - **QA:** Embedding-based personalized retrieval model. It adopts a query-aware user modeling technique using the Attention mechanism to weigh the contribution of the user-related documents in composing the user representation.
  - **BiEnc + Mean:** Weighted sum-based fusion of BiEnc and Mean.
  - **BiEnc + QA:** Weighted sum-based fusion of BiEnc and QA.

We aggregated the document scores of each re-rankers the original BM25 scores using the weighted sum fusion algorithm implemented in `ranx.fuse`.

Results

Table 1. Effectiveness of the compared models. † denote significant improvements in a Bonferroni corrected Two-sided Paired Student's t-Test with $p < 0.001$ over all the baselines. Best results are highlighted in boldface.

Model	Computer Science			Physics			Political Science			Psychology		
	MAP	MRR	NDCG	MAP	MRR	NDCG	MAP	MRR	NDCG	MAP	MRR	NDCG
BM25	12.25	48.93	22.45	12.77	53.68	26.88	13.27	50.23	24.07	12.58	51.19	23.93
BM25 + Pop	16.64	58.63	28.97	15.45	59.61	30.88	16.04	57.39	28.46	15.13	56.46	27.29
BM25 + BiEnc	18.21	58.02	28.90	16.98	60.99	32.13	18.15	57.64	29.25	20.72	63.41	33.11
BM25 + All-BiEnc	17.82	57.79	28.59	16.81	60.95	32.02	18.51	58.82	29.94	20.23	62.91	32.64
BM25 + Mean	16.37	54.57	26.69	16.44	59.05	31.18	16.61	55.05	27.58	16.73	57.05	28.42
BM25 + QA	17.88	57.21	28.49	17.47	61.80	32.72	17.69	57.58	28.94	18.90	60.92	31.12
BM25 + BiEnc + Mean	19.92	60.64	30.80	18.91	63.88	34.51	19.26	59.76	30.63	21.97	65.21	34.71
BM25 + BiEnc + QA	**20.11†**	**61.17**	**31.15†**	**18.98**	**64.78**	**34.81**	**19.85†**	**61.20†**	**31.41†**	**21.99**	**65.62**	**34.85**

Online Resources

- To learn more about our *Multi-Domain Benchmark for Personalized Search Evaluation* scan the QR Code below: