Supporting Conference Organizers in Expert-Finding Tasks using Transformer-based Embedding and Retrieval

Tom Eichhorn[1,*], Benedict Hartmann[1] and Nico Lindemann[1]

[1]University in Hagen, Universitätsstr. 1, 58097 Hagen, Germany

Abstract

Expert finding (EF) is a central task in scientific conference organization, particularly for reviewer assignment and program planning, yet remains challenging due to increasing submission volumes and heterogeneous metadata. While transformer-based embeddings enable semantic representation of scientific texts, there is a lack of reproducible data preparation pipelines and cloud-capable retrieval frameworks that integrate abstracts, author information, and semantic search in conference management contexts. This position paper argues that these two gaps can be addressed through a modular, transformer-based architecture for supporting Professional Congress Organizers (PCOs) in EF tasks. We propose (i) a reproducible embedding pipeline that prepares scientific abstracts and aggregates them into author-level representations, and (ii) a cloud-capable retrieval architecture that combines vector-based semantic search with structured metadata integration. Both designs are grounded in the Nunamaker et al. research methodology and developed through systematic observation of the state of the art and theory building. We further argue that the modular, decoupled nature of the proposed architecture lowers the adoption barrier for PCOs, positioning semantic EF as a practical augmentation of existing conference management workflows rather than a disruptive replacement.

Keywords

Professional Conference Organization, Vector Embedding, Vector Retrieval, Expert Finding

1. Introduction

The **Meetings, Incentives, Conferences, Events (MICE)** industry encompasses conferences, incentive travel, and exhibitions, representing a particularly knowledge-intensive and organizationally complex sector with a global market cap of $645.7 billion in 2022 [1]. In recent years, the COVID-19 pandemic forced rapid adoption of virtual and hybrid event formats [2], climate concerns have driven demands for more sustainable conference practices [3], and advances in **Artificial Intelligence (AI)**, particularly transformer architectures [4], have opened new possibilities for intelligent support systems in conference management, a need equally recognized in data-intensive scientific domains such as genomic diagnostics [5, 6].

Scientific conferences serve as crucial venues for knowledge exchange in academia and Information Technology (IT) [7, 8]. Their organization involves complex coordination among **Professional Congress Organizerss (PCOs)**, authors, reviewers, chairs, and panelists. Among these tasks, **Expert Finding (EF)** is particularly challenging: PCOs must identify and assign appropriate reviewers to submitted papers, select qualified panelists, and recommend domain experts for various roles [9, 10]. This problem has become increasingly critical due to growing submission volumes and the complexity of accurately matching reviewer expertise to paper topics [11, 12].

Traditional approaches rely on manual processes, keyword matching, or simple citation metrics [11, 12]. While some **Conference Management Systemss (CMSs)** offer basic filtering capabilities, they generally lack sophisticated semantic understanding of research topics and cannot effectively capture the nuanced relationships between expert profiles and submission content [10]. A reproducible data

Proceedings of TRI-IVIS 2026, the First International Workshop on Trustable, Reproducible, and Intelligent Information Visualization Systems, June 2026, Venice, Italy.

*Corresponding author.

tom.eichhorn@studium.fernuni-hagen.de (T. Eichhorn); benedict.hartmann@fernuni-hagen.de (B. Hartmann); nico.lindemann@fernuni-hagen.de (N. Lindemann)

0009-0009-6862-3160 (T. Eichhorn); 0000-0001-8932-5853 (B. Hartmann); 0009-0006-8232-6240 (N. Lindemann)

preparation pipeline that systematically transforms conference abstracts and author metadata into semantic vector representations could enable PCOs to leverage modern language models for EF without requiring deep technical expertise [13, 14]. Similarly, a cloud-capable retrieval framework providing intuitive interfaces could make semantic EF practically accessible for conference organizers [6, 5]. Furthermore, systematic evaluation combining retrieval effectiveness with usability assessment would demonstrate its practical viability [15, 16].

Previous work has addressed related challenges: the **Cloud-based Information Extraction (CIE)** project developed cloud-capable NER systems for information extraction [17], and **FIT4NER** advanced entity recognition across domains [18, 14]. The **KGen4MICE** initiative targets MICE challenges including data-driven expert recommendation. Despite these advances, two fundamental gaps remain. First, there is no broadly established, openly documented pipeline integrating conference abstracts and author metadata into a coherent semantic embedding workflow, leaving it unclear how PCOs can concretely transform their existing databases into semantically searchable representations. Second, there is a lack of production-ready retrieval frameworks combining semantic vector representations with structured metadata, cloud deployment, and systematic evaluation of both effectiveness and usability in conference management scenarios.

We structure our research using the Nunamaker et al. framework [19], which interleaves four phases: Observation, Theory Building, Systems Development, and Experimentation. The present paper covers the first two phases and is organized around two research questions:

Research Question 1 (RQ1): How can scientific abstracts and author information be prepared and represented to enable semantic vector embedding for thematic author retrieval?

Research Question 2 (RQ2): How can a cloud-capable retrieval system be developed that suggests relevant authors for thematic queries, independent of data origin?

The two research questions map directly onto the two contributions of this paper. RQ1 motivates the design of a modular embedding pipeline that processes conference abstracts and aggregates them into author-level representations. RQ2 motivates the design of a cloud-capable retrieval architecture that combines semantic vector search with structured metadata. Both designs are grounded in the state of the art surveyed in Section 2 and operationalized in the conceptual architecture presented in Section 3.

2. State of the Art in Science and Technology

Contextual embeddings produced by transformer models [20, 21] have become the de-facto standard for semantic text representation, superseding sparse approaches such as TF-IDF and BM25 [22]. Models fine-tuned for sentence or document similarity, such as Sentence-BERT (SBERT) [21], MiniLM [23], and MPNet [24], produce dense representations suitable for large-scale nearest-neighbor retrieval. Domain-specific variants such as SciBERT [25] further improve performance on scientific corpora. For author representation, several studies aggregate publication embeddings into author-level vectors [13, 26] and report strong retrieval performance on bibliographic benchmarks. However, these studies assume homogeneous, preprocessed datasets, but none provides an openly documented, end-to-end pipeline that handles the heterogeneous metadata conditions typical of live CMS environments.

Remaining Challenge 1 (RC1): A reproducible, openly documented pipeline integrating scientific abstracts and author metadata for semantic vector embedding in MICE contexts does not yet exist.

Dense retrieval frameworks such as FAISS [27], Weaviate, and Elasticsearch support approximate nearest-neighbor search at scale and have become standard infrastructure for vector-based Information Retrieval (IR) systems. Neural and dense retrieval models address the semantic gap of traditional term-based approaches by embedding queries and documents into a shared vector space [28], while hybrid methods that combine sparse lexical signals with dense representations further improve robustness across diverse query types [29, 30]. To ensure reusability and scalability, modern retrieval systems increasingly adopt containerized microservice architectures: frameworks like Docker and FastAPI facilitate deployment, isolation, and dynamic scaling of components such as embedding generation,

indexing, and retrieval APIs. Although these frameworks provide strong infrastructural support, they lack domain-specific adaptation for EF in conference contexts. Furthermore, usability has rarely been evaluated alongside retrieval effectiveness in CMS scenarios, leaving it unclear whether technically sound systems are also practically accessible to PCOs.

Remaining Challenge 2 (RC2): A cloud-capable retrieval framework combining semantic vector search with author metadata, systematically evaluated for both retrieval quality and usability in MICE scenarios, is required.

From the outlined state of the art, two design requirements emerge: a system architecture that processes scientific abstracts and author information into consistent semantic representations (RC1), and a cloud-capable retrieval framework with hybrid query processing that is evaluated for both retrieval effectiveness and practical usability (RC2). The following section addresses these requirements through a conceptual system design.

3. Modeling and Design

Following Nunamaker et al. [19], this section establishes the conceptual foundation for the system, operationalizing the embedding pipeline (RQ1) and the retrieval architecture (RQ2). The conceptual modeling approach integrates **User Centered System Design (UCSD)** [31, 32] to ensure methodological rigor and user relevance, placing the needs of PCOs and administrators at the center of design decisions. **Unified Modeling Language (UML)** [33] is used selectively to provide formal notation for describing data structures, interactions, and system components, supporting the theoretical formulation of both research objectives.

Use Cases and Information Model

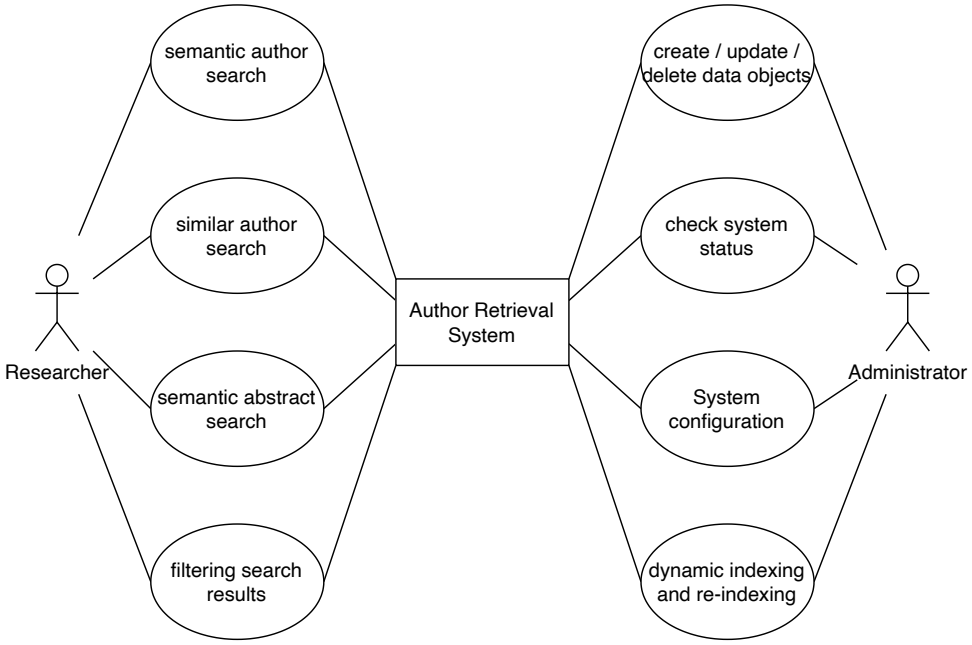

Figure 1: UML Use Case Diagram showing the application use cases.

As illustrated in Figure 1, the system supports two actor groups. *Researchers* interact through four use cases: (UC1) semantic author search for expert finding, (UC2) semantic abstract search, (UC3) similar author search by embedding proximity, and (UC4) metadata-based result filtering. *Administrators* manage the system through (UC5) CRUD operations on data objects, (UC6) dynamic index rebuilding, (UC7) system configuration, and (UC8) health monitoring.

To illustrate how these use cases interact in practice, consider the following scenario. A PCO receives a submission on *privacy-preserving federated learning for clinical NLP*. No obvious reviewer comes to mind, as the topic spans machine learning, privacy, and biomedical text processing. The PCO enters the abstract as a free-text query (UC1). The system encodes the query on-the-fly, retrieves the ten most semantically similar author profiles, and returns a ranked list with titles of associated publications. Noticing that the top result specialises in federated learning but has limited NLP background, the PCO navigates to that author's profile and invokes the similar-author view (UC3), which surfaces two researchers with stronger NLP publication records. A topic filter (UC4) further narrows the candidates to those affiliated with a European institution, satisfying a regional balance requirement. The entire process should take less than a minute and requires no knowledge of the underlying vector indices or embedding models.

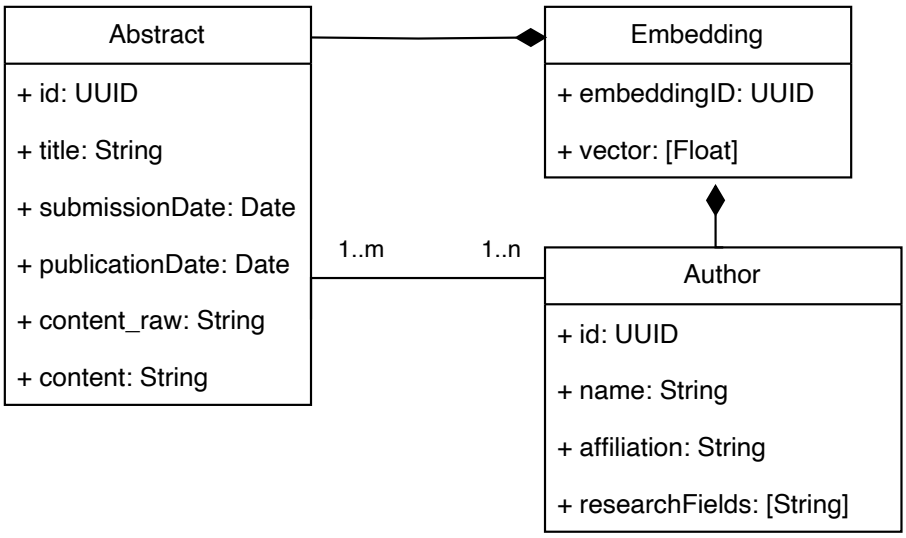

Figure 2: UML Class Diagram showing the relationship between experts, documents, and topics.

The information model comprises three core entity classes. `Abstract` stores the submission with identifier, title, submission and publication dates, and both raw and cleaned textual content, preserving the original text alongside a normalised version suitable for embedding. `Author` captures the expert profile with identifier, name, institutional affiliation, and a list of research fields, providing the structured metadata needed for filtering and result enrichment. `Embedding` holds the dense vector representation linked to both abstracts and authors via a many-to-many relationship, identified by an embedding ID and a float vector array. This many-to-many design reflects the reality that a single author is associated with multiple abstracts, and that each abstract may be embedded under different model configurations for comparative purposes. The deliberate separation of `Embedding` from the domain entities follows established practices in modular retrieval system design [27]: vector representations can be regenerated with improved models without affecting metadata integrity, and multiple embedding configurations can coexist for evaluation or model comparison.

While the information model defines *what* data is stored and how entities relate, the system architecture defines *how* this data is processed, indexed, and made retrievable at runtime.

System Architecture

Figure 3 shows the component architecture mapped to four functional layers, each with clearly separated responsibilities. The *Persistence layer* forms the foundation of the system, separating relational metadata storage (PostgreSQL) from vector-based similarity search (FAISS [27]). This decoupling allows each storage backend to be scaled, replaced, or upgraded independently. The *Core layer* builds upon persistence by providing three services: the Embedding component handles transformer-based

vector generation for abstracts and authors; the Index Manager controls the lifecycle of FAISS indices including incremental updates and rebuilds; and the Data Import component validates and ingests structured metadata. The *Search layer* exposes three domain-specific retrieval services: Author Search for expert finding by semantic query, Abstract Search for document-level retrieval, and Similar Author for embedding-proximity-based recommendations. Finally, the *API layer* exposes a single Search Orchestrator that acts as the unified gateway, routing incoming requests to the appropriate search service and coordinating response aggregation. This layered design ensures that components can be developed, deployed, and scaled independently [6].

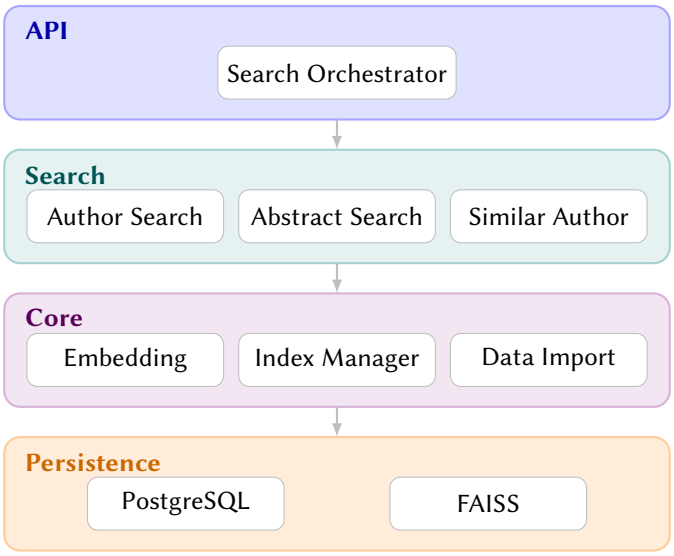

Figure 3: Four-layer component architecture with inter-layer communication. Components are grouped by functional responsibility, with the Persistence layer being storage-technology agnostic.

Embedding and Retrieval Workflow

The embedding workflow addresses RQ1 by ingesting structured JSON containing abstracts and author profiles. After validation and deduplication, textual content is encoded by a configurable transformer model from the SBERT family [21], such as `all-MiniLM-L6-v2` [23] or domain-adapted variants like SciBERT [25] for scientific corpora. Each abstract is encoded into a dense vector in a shared semantic space. Author-level embeddings are then constructed by mean-aggregating the embeddings of all abstracts associated with a given author, a strategy shown to be effective for academic expert representation [13, 34]. The design decouples metadata management from embedding generation: new submissions trigger incremental embedding updates without requiring full index rebuilds, and improved language models can be substituted without altering the underlying relational schema.

The retrieval workflow addresses RQ2 through a dual-path query strategy. For *semantic queries*, the user's free-text input is encoded on-the-fly using the same transformer model, and the resulting query vector is matched against precomputed author or abstract embeddings via an oversampled k-nearest-neighbour search in the FAISS index [27]. Oversampling compensates for candidates subsequently removed by metadata filters, preserving result quality in paginated views. Results are then enriched with structured metadata from the relational database before being returned to the user. For *metadata-only queries*, the embedding step is bypassed entirely and the database is queried directly, reducing latency for filtered lookups that do not require semantic ranking. This hybrid strategy draws on insights from large-scale IR benchmarking [30] and provides natural extension points for future improvements such as cross-encoder re-ranking [29] or lexical boosting via BM25.

4. Summary and Future Work

This paper has argued for a transformer-based approach to semantic EF in the MICE and conference management domain. Grounded in the Nunamaker et al. methodology [19], we identified two unresolved challenges through observation of the state of the art and addressed them through a conceptual system design. RC1 is resolved by a modular embedding pipeline that transforms heterogeneous conference metadata into author-level vector representations using models such as SBERT [21] and SciBERT [25], following aggregation strategies validated in prior expert retrieval work [13, 34]. RC2 is resolved by a four-layer retrieval architecture that combines FAISS-based vector search [27] with structured metadata, supporting both semantic and metadata-only query paths in the spirit of hybrid retrieval benchmarking [30].

Beyond technical feasibility, we argue that semantic EF represents a qualitative shift in how PCOs interact with scientific knowledge. Current practice reduces a researcher's publication record to keywords or h-index values, discarding the semantic relationships that make expert matching meaningful. Dense vector representations [20, 21] restore this richness by capturing thematic proximity across terminological variation between subfields. For the MICE industry, where reviewers are recruited across institutional and disciplinary boundaries [11], this directly affects the fairness and scientific value of the review process. The modular, cloud-native design further lowers the adoption barrier for PCOs who lack dedicated IT infrastructure: by decoupling embedding generation from retrieval and both from the underlying CMS, the architecture integrates incrementally into existing workflows [6, 5]. Ensuring that the resulting interfaces remain comprehensible and free of usability deficiencies is equally essential [15, 16].

Future work will operationalize the proposed design through a prototypical implementation and systematic evaluation against a gold standard benchmark, covering intrinsic embedding quality and established IR retrieval metrics. Integration with heterogeneous CMSs through wrapper-mediator architectures and extension toward hybrid lexical-semantic retrieval [29] represent the primary directions for subsequent research.

Declaration on Generative AI

During the preparation of this work, the authors used Claude's Sonnet 4.6 in order to: Grammar and spelling check. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Custom Market Insights, Global mice market report - 2023 to 2032, 2024. URL: https://www.custommarketinsights.com/report/mice-market/, accessed: 2024-09-03.

[2] V. Pavluković, M. Vujičić, J. Kennell, M. Cimbaljević, Sustainability in the meetings industry, in: Proceedings of the Singidunum International Tourism Conference - Sitcon 2020, Singidunum University, 2020, pp. 11–17. URL: https://singipedia.singidunum.ac.rs/izdanje/43286-sustainability-in-the-meetings-industry. doi:10.15308/Sitcon-2020-11-17.

[3] Y. Yuan, J. Enerio, W. Zhang, A guide to sustainable events: Singapore mice carbon calculator (2024).

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[5] T. Krause, P. Tamla, A. Leoni, F. Monti, F. De Luzi, J. F. Gerald, B. Andrade, H. Afli, M. Mecella, P. Buono, A. Molinari, M. Hemmje, From reads to reports: A vision for a gfm-powered genomic diagnostic platform, in: 2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2025, pp. 7207–7213. doi:10.1109/BIBM66473.2025.11356528.

[6] P. Tamla, T. Krause, M. Hemmje, F. Monti, F. De Luzi, M. Mecella, B. Andrade, P. Buono, A. Molinari, The gendai cloud-native infrastructure and data stewardship for clinical metagenomic diagnostics, in: 2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2025, pp. 7229–7236. doi:10.1109/BIBM66473.2025.11356811.

[7] B. Fowler, P. W. Cardon, B. Marshall, K. Elder, What do conference attendees want from academic presentations? a study of an information systems professional organization., Issues in Information Systems 22 (2021).

[8] D. Severt, Y. Wang, P.-J. Chen, D. Breiter, Examining the motivation, perceived performance, and behavioral intentions of convention attendees: Evidence from a regional conference, Tourism management 28 (2007) 399–408.

[9] M. Stankovic, C. Wagner, J. Jovanovic, P. Laublet, Looking for experts? what can linked data do for you?, in: LDOW, 2010.

[10] O. Husain, N. Salim, R. A. Alias, S. Abdelsalam, A. Hassan, Expert finding systems: A systematic review, Applied Sciences 9 (2019) 4250.

[11] J. Jovanovic, E. Bagheri, Reviewer assignment problem: A scoping review, 2023. URL: http://arxiv.org/abs/2305.07887. doi:10.48550/arXiv.2305.07887. arXiv:2305.07887 [cs].

[12] A. C. Ribeiro, A. Sizo, L. P. Reis, Investigating the reviewer assignment problem: A systematic literature review (2023) 01655515231176668. URL: https://journals.sagepub.com/doi/full/10.1177/01655515231176668. doi:10.1177/01655515231176668, publisher: SAGE Publications Ltd.

[13] M. Berger, J. Zavrel, P. Groth, Effective distributed representations for academic expert search, in: Proceedings of the First Workshop on Scholarly Document Processing, 2020, pp. 56–71. URL: http://arxiv.org/abs/2010.08269. doi:10.18653/v1/2020.sdp-1.7. arXiv:2010.08269 [cs].

[14] P. Tamla, F. Freund, M. Hemmje, Cloud-based medical named entity recognition: A fit4ner-based approach, Information 16 (2025) 395. doi:10.3390/info16050395.

[15] P. Buono, D. Caivano, M. F. Costabile, G. Desolda, R. Lanzilotti, Towards the detection of ux smells: The support of visualizations, IEEE Access 8 (2020) 6901–6914. doi:10.1109/ACCESS.2019.2961768.

[16] P. Buono, N. Berthouze, M. F. Costabile, A. Grando, A. Holzinger, Special issue on human-centered artificial intelligence for one health, Artificial Intelligence in Medicine 156 (2024) 102946. doi:https://doi.org/10.1016/j.artmed.2024.102946.

[17] P. Tamla, B. Hartmann, N. Nguyen, C. Kramer, F. Freund, M. Hemmje, Cie: A cloud-based information extraction system for named entity recognition in aws, azure, and medical domain, in: F. Coenen, A. Fred, D. Aveiro, J. Dietz, J. Bernardino, E. Masciari, J. Filipe (Eds.), Knowledge

Discovery, Knowledge Engineering and Knowledge Management, Springer Nature Switzerland, Cham, 2023, pp. 127–148.
[18] F. Freund, P. Tamla, T. Reis, M. Hemmje, P. M. Kevitt, FIT4NER - Towards a Framework-Independent Toolkit for Named Entity Recognition, 2023. 8th Collaborative European Research Conference (CERC 2023), Barcelona, Spain, June 9-10, 2023.
[19] J. F. Nunamaker, M. Chen, T. D. M. Purdin, Systems development in information systems research 7 (1990) 89–106. URL: http://www.jstor.org/stable/40397957, publisher: Taylor & Francis, Ltd.
[20] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423/.
[21] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-networks, 2019. URL: http://arxiv.org/abs/1908.10084. doi:10.48550/arXiv.1908.10084. arXiv:1908.10084 [cs].
[22] S. Robertson, H. Zaragoza, The probabilistic relevance framework: BM25 and beyond 3 (2009) 333–389. URL: http://www.nowpublishers.com/article/Details/INR-019. doi:10.1561/1500000019.
[23] 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: Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 5776–5788. URL: https://proceedings.neurips.cc/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.
[24] K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu, MPNet: Masked and permuted pre-training for language understanding, in: Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 16857–16867. URL: https://proceedings.neurips.cc/paper/2020/hash/c3a690be93aa602ee2dc0ccab5b7b67e-Abstract.html.
[25] I. Beltagy, K. Lo, A. Cohan, SciBERT: A pretrained language model for scientific text, 2019. URL: https://arxiv.org/abs/1903.10676v3.
[26] R. Brochier, A. Gourru, A. Guille, J. Velcin, New datasets and a benchmark of document network embedding methods for scientific expert finding, 2020. URL: http://arxiv.org/abs/2004.03621. doi:10.48550/arXiv.2004.03621. arXiv:2004.03621 [cs].
[27] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with GPUs 7 (2021) 535–547. URL: https://ieeexplore.ieee.org/abstract/document/8733051. doi:10.1109/TBDATA.2019.2921572.
[28] V. Karpukhin, B. Oguz, S. Min, P. S. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. Yih, Dense passage retrieval for open-domain question answering., in: EMNLP (1), 2020, pp. 6769–6781.
[29] O. Khattab, M. Zaharia, ColBERT: Efficient and effective passage search via contextualized late interaction over BERT, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '20, Association for Computing Machinery, 2020, pp. 39–48. URL: https://dl.acm.org/doi/10.1145/3397271.3401075.
[30] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, I. Gurevych, BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models, 2021. URL: http://arxiv.org/abs/2104.08663. doi:10.48550/arXiv.2104.08663. arXiv:2104.08663 [cs].
[31] D. A. Norman, S. W. Draper, User centered system design: new perspectives on human-computer interaction, L. Erlbaum, 1986.
[32] J. Gulliksen, B. Göransson, I. Boivie, S. Blomkvist, J. Persson, Å. Cajander, Key principles for user-centred systems design 22 (2003) 397–409. URL: http://www.tandfonline.com/doi/abs/10.1080/01449290310001624329. doi:10.1080/01449290310001624329.
[33] G. Booch, J. Rumbaugh, I. Jacobson, The unified modeling language user guide, The Addison-Wesley object technology series, Addison-Wesley, 1999.
[34] H. Liu, Z. Lv, Q. Yang, D. Xu, Q. Peng, Expertbert: Pretraining expert finding, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, CIKM '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 4244–4248. URL: https://doi.org/10.1145/3511808.3557597. doi:10.1145/3511808.3557597.