Content-Based Dense Retrieval of Open Datasets

Qiaosheng Chen[1]

[1]State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China

Abstract

The rapid growth of open data has intensified the need for effective dataset search capabilities. This research proposal focuses on enhancing dataset search through content-based dense retrieval, addressing the limitations of current metadata-dependent systems. This research aims to tackle the challenges of dataset size, heterogeneity, and the creation of a comprehensive test collection for evaluation. The proposed research methods include data summarization techniques for large datasets and a unified representation of heterogeneous data, which are inspired by research related to the Semantic Web. Additionally, the research will explore a coarse-to-fine tuning strategy for dense retrieval models, leveraging data augmentation through distant supervision and self-training. The evaluation plan involves constructing a content-based test collection and comparing retrieval performance between metadata-only and content-enhanced approaches. The expected outcome is the development of effective content-based dataset search solutions, ultimately improving data findability.

Keywords

Dataset Search, Dense Retrieval, Open Data

1. Introduction

The availability and significance of open data have led to a surge in interest and reliance on dataset search within the field of information retrieval [1]. However, represented by Google Dataset Search [2], existing approaches and systems predominantly rely on metadata (descriptive text for dataset, such as title, description), which often suffers from low quality and limited availability. These metadata-based approaches have posed shortage in accurately capturing the relevance of datasets. For addressing the gap between users' real data needs and the quality of dataset metadata, it necessitates a shift towards content-based approaches that can effectively harness the richness of dataset content [3, 4].

On the other hand, dense retrieval models, which have become mainstream in the field of document retrieval [5], have not yet been fully explored in the field of dataset search. In particular, how to apply dense retrieval models to content-based dataset search problems still faces many challenges. First, the large size of dataset content poses computational challenges, especially when it exceeds the processing capacity of standard dense retrieval models which are mainly based on pre-trained language models (PLMs). Additionally, the heterogeneity of dataset content, spanning various data formats and domains [6], further complicates the development of unified content-based search solutions.

Proceedings of the Doctoral Consortium at ISWC 2024, co-located with the 23rd International Semantic Web Conference (ISWC 2024)

qschen@smail.nju.edu.cn (Q. Chen)
https://cqsss.github.io (Q. Chen)
0009-0002-0610-7725 (Q. Chen)

The proposed research aims to contribute towards the development of robust and effective content-based solutions for dataset search, ultimately improving the findability and reusability of open datasets. Users across various domains, including researchers, data scientists, policymakers, and businesses, will benefit from content-based dataset search, while professional researchers in fields such as information retrieval, natural language processing (NLP), and machine learning are particularly invested in its advancement. Industries relying heavily on data-driven decision-making, such as healthcare, finance, agriculture, and environmental science, should also care about its development. Beyond the domain of information retrieval, this research involves technologies relevant to the Semantic Web and Knowledge Graph (KG). RDF datasets represent a significant part of open data. Moreover, employing ontologies or KGs as a framework can aid in analyzing the content of open datasets and processing heterogeneou data from a unified perspective. The advancement of dataset search also stands to catalyze the realization of findable, accessible, interoperable, and reusable (FAIR) open data within the Semantic Web community.

2. Related Work

In this section, we review recent advancements in dataset search and dense retrieval, highlighting limitations of current dataset search methods and examining strengths of dense retrieval techniques.

2.1. Dataset Search

Dataset search has garnered increasing attention with the proliferation of diverse and voluminous datasets, prompting the development of search approaches and systems [1, 7]. Notably, Google Dataset Search [2] has paved the way as a pioneering dataset search engine, enabling keyword retrieval over published metadata of Web datasets. However, its reliance on metadata limits its effectiveness in supporting queries oriented towards dataset content. Moreover, existing dataset retrieval test collections [8, 9, 10] primarily depend on metadata annotations during construction, resulting in a lack of evaluation benchmarks for content-based dataset search.

Recent studies have highlighted the importance of integrating considerations for dataset content to enhance search effectiveness. Ota et al. [11] utilized value co-occurrence information within tabular datasets to infer attribute domains, while Chen et al. [12] proposed a BERT-based ranking model for table retrieval, focusing on selecting the most salient table items as representatives of the entire dataset. StruBERT [13] introduced a structure-aware BERT model to capture both structural and textual information of tabular datasets. Moreover, existing tabular dataset or RDF dataset search systems such as Auctus [14], LODAtlas [15], and CKGSE [16] leverage dataset content to augment retrieval capabilities and enhance user search experiences. *However, these efforts primarily focus on single-format data, such as tabular or RDF data, overlooking the challenges posed by multi-format datasets.*

2.2. Dense Retrieval

Recent advancements in dense retrieval have been significantly influenced by the incorporation of PLMs, which have demonstrated remarkable capabilities in capturing semantic nuances

within text [5]. This approach, often referred to as dense retrieval, leverages the dense vector representations (embeddings) of text to facilitate semantic matching between queries and documents. Notably, Karpukhin et al. [17] presented dense passage retrieval (DPR) for open-domain question answering, highlighting the effectiveness of PLMs in this context. Their work has been seminal in shaping subsequent research. The concept of using multiple representations for improved text encoding has been explored by Humeau et al. [18] through their poly-encoder architectures, which allow for richer semantic interactions between queries and texts. The challenge of training efficient and robust dense retrievers has been addressed by various work. For instance, Gao and Callan [19] introduced Condenser, a pre-training architecture specifically designed to improve dense retrieval. Nogueira et al. [20] demonstrated the effectiveness of multi-stage document ranking using BERT, showcasing how PLMs can be effectively integrated into reranking stage. Furthermore, ColBERT [21] has provided insights into efficient and effective passage search through contextualized late interaction over BERT. *Most of the current dense retrieval methods focus on retrieval of text documents or passages, whereas the structured content of datasets requires new dense model structures or retrieval strategies. The large data size and complex heterogeneity also make it difficult to directly treat the dataset content as plain text.*

3. Problem Statement

In this section, we discuss the typical composition of datasets, outline the problem of content-based dataset search, and formulate hypotheses and research questions derived from our investigation.

To clarify the distinction between dataset search and general document search, we first introduce the composition of a dataset, which consists of the following two parts:

- *Metadata*: This part includes descriptive fields provided by the dataset publisher, such as title, description, publishing organization, and other useful information about the dataset.
- *Data Files*: A dataset consists of various data files, potentially in different formats. This research only considers textual data files, including unstructured TXT, PDF, and DOC files, as well as structured files such as graph data (RDF, OWL), tabular data (CSV, XLS), and key-value pair data (JSON, XML). Images (JPEG, PNG), videos (AVI, MPG), audios (WAV, MP3), and other non-textual formats are excluded from the scope of this research.

The research focuses on *ad hoc dataset retrieval* [8, 3], the foundational form of dataset search. This process involves retrieving, from a collection D of datasets, a ranked list of datasets $\langle d_1, d_2, \ldots \rangle$ that are most relevant to a keyword query q. The relevance assessment between query q and each dataset $d \in D$ is conducted independently of other datasets $d' \in D$, where $d \neq d'$. The primary objective is to compute the relevance score of each dataset $d \in D$ to a given keyword query q. The prevalent dense retrieval paradigm typically employs a PLM as an encoder $E(\cdot)$ to encode a dataset d and a query q into vectors $\mathbf{v}_d$ and $\mathbf{v}_q$ respectively. Subsequently, it computes the similarity score between these vectors to gauge the relevance of d to q.

According to studies on metadata quality [22, 23], the metadata of open datasets on the Web often lacks guaranteed quality and is underutilized by both publishers and users. Meanwhile,

dense retrieval models based on PLMs have exhibited increasingly powerful text understanding capabilities with advancements in NLP [5]. Hence, the application of dense retrieval models in dataset search becomes imperative. Based on these findings, we propose the following main hypothesis and research question:

Hypothesis. Dataset metadata quality often varies and may not fully describe the content. Users frequently seek information from the actual data files, and content-focused queries may not align well with the available metadata.

RQ0. To what extent can content-based dense dataset retrieval methods outperform traditional metadata-centered approaches?

Building upon the hypothesis and RQ0, this research investigates the application of dense models to content-based dataset retrieval. Nonetheless, representing and indexing complex dataset content with PLM-based dense models presents substantial challenges. To address these challenges, we decompose RQ0 into the following four specific research questions:

RQ1. How to overcome the challenge presented by the extensive size of dataset content, especially when it exceeds the processing capacity of dense retrieval models?

RQ2. How to address the heterogeneity of dataset content, encompassing variations in formats?

RQ3. How to develop a dataset retrieval test collection which considers the content of datasets, rather than annotated solely relies on metadata?

RQ4. How to enhance the size and quality of existing public dataset retrieval test collections, particularly in terms of providing sufficient training data for dense retrieval models?

4. Research Methods

In this section, we provide a detailed and systematic research methodology that outlines how we address each research question (RQ1-RQ4) and validate our hypotheses [3, 24, 25, 26]. The methodology is structured to ensure a comprehensive and coherent approach to solving the challenges of content-based dense retrieval of open datasets.

RQ1. To overcome the challenge posed by the large size of dataset content that exceeds the input capacity of PLMs, this research proposed an approach involving the extraction of a data summary for each dataset. Starting with RDF datasets, we introduced a technique to handle large RDF datasets by extracting a compact, representative subset of RDF triples [25]. This subset was selected to preserve the semantic integrity of the dataset and was used to create a document representation that fits within the token limit of dense ranking models. We employed two of the existing static RDF dataset summarization methods, IlluSnip [27, 28] selecting top-ranked RDF triples covering the most frequent classes, properties, and entities, and PCSG [29] extracting a connected subgraph from an RDF graph covering as many data patterns as possible. Furthermore, we proposed a dynamic data summary extraction method for dataset search, selecting compact data snippets of appropriate size that are relevant to the user query [26]. By integrating these methods, one can create a compact, semantically representative, and query-biased data summary of the original dataset. This enables the use of PLMs for tasks such as dense dataset retrieval, where the models can process the summarized data to understand and rank datasets based on their relevance to user queries without being

hindered by size limitations.

RQ2. We address the challenge of heterogeneity in dataset content by transforming data from various formats into a unified representation. The method establishes mapping rules for structured data, such as graph data, tabular data, and key-value pair data. These rules convert the heterogeneous data into unified data chunks. Each data chunk is modeled as a set of data triples, which consist of a subject, a predicate, and an object. This triple-structured format allows for uniform processing of all datasets, regardless of their original format. Converting different data formats into unified data chunks creates a consistent input for dense ranking models. This approach allows for the exploitation of heterogeneous data in dataset ranking, overcoming the limitations imposed by the diverse formats of open data. The summarized data chunks can then be used to rank datasets based on their relevance to a given query, thus enhancing the search accuracy and making the process more efficient. To ensure that the structural information is not lost during the conversion process, the mapping rules we employed preserve the hierarchical and relational aspects of the original data. For graph data, we maintain the relationships between nodes and edges by representing them as triples. For tabular data, we preserve the row-column structure by mapping rows to subjects and columns to predicates. For key-value pair data, we maintain the key-value relationships by representing keys as predicates and values as objects. This approach ensures that the structural integrity of the original data is preserved, which is beneficial for accurate retrieval. Additionally, we conduct experiments to evaluate the impact of content on retrieval performance, providing insights into the importance of preserving this information during the conversion process.

RQ3. We released a content-based RDF dataset retrieval test collection ACORDAR [3], and subsequently enhanced it to build ACORDAR 2.0 [24]. Constructing this content-based dataset retrieval test collection began with the collection of RDF datasets from various open data portals, ensuring a diverse and representative sample. Keyword queries were then formulated, either by analyzing user needs or through crowd-sourcing, resulting in a set of search terms that reflected actual information demands. To accommodate the complexity and size of datasets, a dashboard was developed to assist annotators in browsing and understanding the content of datasets. This tool was crucial for creating content-oriented queries and making informed relevance judgments. Annotators used the dashboard to analyze datasets and generate queries that capture the dataset's essence. These queries were then used to pool potentially relevant datasets, which were subsequently annotated for relevance. The pooling process was done using both sparse and dense retrieval models to ensure a broad coverage of potential matches. Relevance judgments were made on a graded scale, with annotators assessing the degree to which each dataset met the query's requirements. To ensure quality, annotations involved multiple annotators and a validation process. ACORDAR 2.0 was further enriched by transforming keyword queries into question-style queries using a large language model (LLM), which increased the diversity of the queries and simulates more natural information-seeking behavior. Our test collection provides a benchmark for evaluating content-based dataset retrieval systems.

RQ4. To address the challenge of limited large labeled datasets necessary for training dense retrieval models, we proposed a coarse-to-fine tuning strategy [25]. This strategy involved an initial coarse-tuning phase with weak supervision obtained from a large set of automatically generated queries and relevance labels. It incorporated two data augmentation methods: distant supervision and self-training. In the distant supervision method, the title of each dataset served

as a query, and the metadata document was assumed to be relevant to this query, thereby generating numerous labeled examples. Meanwhile, the self-training method employed dataset-to-query generators trained on labeled data to generate queries from unlabeled data, further expanding the datasets for training dense models.

This systematic methodology ensures that each research question is addressed with a clear and structured approach, leading to the validation of our hypotheses and the development of effective content-based dataset search solutions.

5. Evaluation

The evaluation plan for this research involves constructing content-based dataset retrieval test collections [3, 24] following the methodology outlined in Section 4. Dataset retrieval and reranking experiments will be conducted on these test collections, as well as on existing public dataset retrieval test collections [8]. Performance will be assessed using commonly used information retrieval metrics such as Recall, Normalized Discounted Cumulative Gain (NDCG), and Mean Average Precision (MAP). The primary objectives of these experiments are as follows:

1. To compare the retrieval performance using solely metadata against retrieval using metadata combined with content.
2. To assess the performance disparity between dense retrieval models and traditional sparse retrieval models in the dataset search scenario.
3. To analyze the impact of various data summarization methods for representing data content in dataset retrieval.
4. To investigate the effectiveness of different query types and characteristics in both metadata-based and content-based retrieval methods.

Comparison of Metadata-Only vs. Content-Enhanced Retrieval. We will conduct experiments to compare the retrieval performance of systems that use only metadata against those that combine metadata with content. This analysis will assess the extent to which content-based retrieval improves search accuracy and relevance. We will examine performance metrics across various dataset types and query scenarios to identify specific cases where content-based retrieval provides significant advantages.

Performance Disparity Between Dense and Sparse Retrieval Models. We will compare the performance of dense retrieval models, which use PLMs, with traditional sparse retrieval models like BM25, which rely on term frequency-based scoring. This evaluation will highlight the strengths and limitations of dense retrieval models in dataset search. By analyzing their performance across diverse query types and datasets, we aim to identify scenarios where dense models excel, particularly in capturing semantic relevance, versus scenarios where sparse models may be more effective.

Impact of Data Summarization Methods. The role of data summarization methods in improving retrieval performance will be analyzed by testing both static techniques, such as IlluSnip and PCSG, and dynamic methods, which generate query-biased summaries. We will evaluate how these summarization approaches influence the relevance and efficiency of dataset

retrieval. Additionally, we will explore the trade-offs between summarization quality and computational cost, providing insights into balancing performance with resource demands.

Query Type and Characteristic Analysis. A detailed examination of different query types and their characteristics will be conducted to understand their effectiveness in metadata-based and content-based retrieval methods. We hypothesize that specific queries requiring detailed content comprehension or precision may benefit more from content-based retrieval. On the other hand, more general queries or those that can be effectively addressed with metadata alone may exhibit similar performance across both approaches. This analysis will help refine retrieval strategies based on query requirements.

In addition, given that the eventual deployment of this work is envisioned in real-world dataset search applications, it is imperative to evaluate the time efficiency and additional space requirements of modules such as data summarization and dense retrieval models when processing real-world open datasets.

6. Conclusion and Future Work

The research proposal on content-based dense retrieval of open datasets is crucial in navigating the vast landscape of available data resources. By shifting the focus from metadata to the actual content of datasets, we can enhance search accuracy, ultimately facilitating more informed decision-making in data discovery. The long-term value of this research lies in its potential to streamline access to diverse datasets, empowering researchers, businesses, and policymakers with valuable insights.

6.1. Limitations and Challenges

Content-based dataset retrieval systems face several limitations and challenges. Time efficiency is a critical issue, as dense retrieval models and summarization techniques require significant computational resources, particularly when processing large datasets. Storage requirements are another concern, as pre-trained language models and their embeddings demand substantial space, making deployment difficult in resource-constrained environments. The heterogeneity and complexity of dataset formats further complicate retrieval, as it is challenging to develop unified solutions that preserve both structural and semantic information. Evaluation is also problematic, as constructing comprehensive and realistic test collections that reflect real-world scenarios is complex yet crucial for assessing system performance. Finally, query understanding remains a persistent challenge, particularly for complex queries that require detailed comprehension of dataset content to map them effectively to relevant datasets.

6.2. Future Research Directions

Future research will focus on several directions to overcome these challenges and enhance dataset retrieval systems. Integrating LLMs into dataset search pipelines offers the potential to improve both accuracy and efficiency, with planned evaluations to quantify their impact on performance metrics. Explainable data summarization techniques will be explored to provide transparent insights into the generation of data summaries and the rationale behind dataset

rankings, fostering trust and usability. Methods for content pattern analysis will be developed to identify and utilize patterns within dataset content, improving retrieval accuracy. Expanding the scope to multi-modal retrieval will address the need to handle diverse data types, including images, videos, and audio, efficiently and at scale. Additionally, real-world deployment of these systems will be prioritized to evaluate scalability and gather user feedback, guiding further refinements and optimizations.

Acknowledgments

The author would like to express his thanks to his supervisor Prof. Gong Cheng for providing helpful suggestions and comments. This work was supported by the NSFC (62072224).

References

[1] A. Chapman, E. Simperl, L. Koesten, G. Konstantinidis, L. Ibáñez, E. Kacprzak, P. Groth, Dataset search: a survey, VLDB J. 29 (2020) 251–272. URL: https://doi.org/10.1007/s00778-019-00564-x. doi:10.1007/S00778-019-00564-X.

[2] D. Brickley, M. Burgess, N. F. Noy, Google dataset search: Building a search engine for datasets in an open web ecosystem, in: The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019, ACM, 2019, pp. 1365–1375. URL: https://doi.org/10.1145/3308558.3313685. doi:10.1145/3308558.3313685.

[3] T. Lin, Q. Chen, G. Cheng, A. Soylu, B. Ell, R. Zhao, Q. Shi, X. Wang, Y. Gu, E. Kharlamov, ACORDAR: A test collection for ad hoc content-based (RDF) dataset retrieval, in: SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022, ACM, 2022, pp. 2981–2991. URL: https://doi.org/10.1145/3477495.3531729. doi:10.1145/3477495.3531729.

[4] J. Chen, X. Wang, G. Cheng, E. Kharlamov, Y. Qu, Towards more usable dataset search: From query characterization to snippet generation, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019, ACM, 2019, pp. 2445–2448. URL: https://doi.org/10.1145/3357384.3358096. doi:10.1145/3357384.3358096.

[5] W. X. Zhao, J. Liu, R. Ren, J. Wen, Dense text retrieval based on pretrained language models: A survey, CoRR abs/2211.14876 (2022). URL: https://doi.org/10.48550/arXiv.2211.14876. doi:10.48550/ARXIV.2211.14876. arXiv:2211.14876.

[6] O. Benjelloun, S. Chen, N. F. Noy, Google dataset search by the numbers, in: The Semantic Web - ISWC 2020 - 19th International Semantic Web Conference, Athens, Greece, November 2-6, 2020, Proceedings, Part II, volume 12507 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 667–682. URL: https://doi.org/10.1007/978-3-030-62466-8_41. doi:10.1007/978-3-030-62466-8_41.

[7] N. W. Paton, J. Chen, Z. Wu, Dataset discovery and exploration: A survey, ACM Comput. Surv. 56 (2023). URL: https://doi.org/10.1145/3626521. doi:10.1145/3626521.

[8] M. P. Kato, H. Ohshima, Y. Liu, H. O. Chen, A test collection for ad-hoc dataset retrieval, in: SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in

Information Retrieval, Virtual Event, Canada, July 11-15, 2021, ACM, 2021, pp. 2450–2456. URL: https://doi.org/10.1145/3404835.3463261. doi:10.1145/3404835.3463261.

[9] T. Cohen, K. Roberts, A. E. Gururaj, X. Chen, S. Pournejati, G. Alter, W. R. Hersh, D. Demner-Fushman, L. Ohno-Machado, H. Xu, A publicly available benchmark for biomedical dataset retrieval: the reference standard for the 2016 biocaddie dataset retrieval challenge, Database J. Biol. Databases Curation 2017 (2017) bax061. URL: https://doi.org/10.1093/database/bax061. doi:10.1093/DATABASE/BAX061.

[10] F. Löffler, A. Schuldt, B. König-Ries, H. Bruelheide, F. Klan, A test collection for dataset retrieval in biodiversity research, Res. Ideas Outcomes 7 (2021) e67887. doi:10.3897/rio.7.e67887.

[11] M. Ota, H. Mueller, J. Freire, D. Srivastava, Data-driven domain discovery for structured datasets, Proc. VLDB Endow. 13 (2020) 953–965. URL: http://www.vldb.org/pvldb/vol13/p953-ota.pdf. doi:10.14778/3384345.3384346.

[12] Z. Chen, M. Trabelsi, J. Heflin, Y. Xu, B. D. Davison, Table search using a deep contextualized language model, in: SIGIR 2020, ACM, 2020, pp. 589–598. URL: https://doi.org/10.1145/3397271.3401044. doi:10.1145/3397271.3401044.

[13] M. Trabelsi, Z. Chen, S. Zhang, B. D. Davison, J. Heflin, Strubert: Structure-aware BERT for table search and matching, in: WWW 2022, ACM, 2022, pp. 442–451. URL: https://doi.org/10.1145/3485447.3511972. doi:10.1145/3485447.3511972.

[14] S. Castelo, R. Rampin, A. S. R. Santos, A. Bessa, F. Chirigati, J. Freire, Auctus: A dataset search engine for data discovery and augmentation, Proc. VLDB Endow. 14 (2021) 2791–2794. URL: http://www.vldb.org/pvldb/vol14/p2791-castelo.pdf. doi:10.14778/3476311.3476346.

[15] E. Pietriga, H. Gözükan, C. Appert, M. Destandau, S. Cebiric, F. Goasdoué, I. Manolescu, Browsing linked data catalogs with lodatlas, in: ISWC 2018, volume 11137 of *Lecture Notes in Computer Science*, Springer, 2018, pp. 137–153. URL: https://doi.org/10.1007/978-3-030-00668-6_9. doi:10.1007/978-3-030-00668-6_9.

[16] X. Wang, T. Lin, W. Luo, G. Cheng, Y. Qu, CKGSE: A prototype search engine for chinese knowledge graphs, Data Intell. 4 (2022) 41–65. URL: https://doi.org/10.1162/dint_a_00118. doi:10.1162/DINT_A_00118.

[17] V. Karpukhin, B. Oguz, S. Min, P. S. H. Lewis, L. Wu, S. Edunov, D. Chen, W. Yih, Dense passage retrieval for open-domain question answering, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, Association for Computational Linguistics, 2020, pp. 6769–6781. URL: https://doi.org/10.18653/v1/2020.emnlp-main.550. doi:10.18653/V1/2020.EMNLP-MAIN.550.

[18] S. Humeau, K. Shuster, M. Lachaux, J. Weston, Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020. URL: https://openreview.net/forum?id=SkxgnnNFvH.

[19] L. Gao, J. Callan, Condenser: a pre-training architecture for dense retrieval, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, Association for Computational Linguistics, 2021, pp. 981–993. URL: https://doi.org/10.18653/v1/2021.emnlp-main.75. doi:10.18653/V1/2021.EMNLP-MAIN.75.

[20] R. F. Nogueira, K. Cho, Passage re-ranking with BERT, CoRR abs/1901.04085 (2019). URL: http://arxiv.org/abs/1901.04085. arXiv:1901.04085.

[21] 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 2020, Virtual Event, China, July 25-30, 2020, ACM, 2020, pp. 39–48. URL: https://doi.org/10.1145/3397271.3401075. doi:10.1145/3397271.3401075.

[22] A. Quarati, Open government data: Usage trends and metadata quality, J. Inf. Sci. 49 (2023) 887–910. URL: https://doi.org/10.1177/01655515211027775. doi:10.1177/01655515211027775.

[23] M. A. Musen, M. J. O'Connor, E. Schultes, M. M. Romero, J. Hardi, J. Graybeal, Modeling community standards for metadata as templates makes data FAIR, CoRR abs/2208.02836 (2022). URL: https://doi.org/10.48550/arXiv.2208.02836. doi:10.48550/ARXIV.2208.02836. arXiv:2208.02836.

[24] Q. Chen, W. Luo, Z. Huang, T. Lin, X. Wang, A. Soylu, B. Ell, B. Zhou, E. Kharlamov, G. Cheng, ACORDAR 2.0: A test collection for ad hoc dataset retrieval with densely pooled datasets and question-style queries, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024, ACM, 2024, pp. 303–312. URL: https://doi.org/10.1145/3626772.3657866. doi:10.1145/3626772.3657866.

[25] Q. Chen, Z. Huang, Z. Zhang, W. Luo, T. Lin, Q. Shi, G. Cheng, Dense re-ranking with weak supervision for RDF dataset search, in: The Semantic Web - ISWC 2023 - 22nd International Semantic Web Conference, Athens, Greece, November 6-10, 2023, Proceedings, Part I, volume 14265 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 23–40. URL: https://doi.org/10.1007/978-3-031-47240-4_2. doi:10.1007/978-3-031-47240-4_2.

[26] Q. Chen, J. Chen, X. Zhou, G. Cheng, Enhancing dataset search with compact data snippets, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024, ACM, 2024, pp. 1093–1103. URL: https://doi.org/10.1145/3626772.3657837. doi:10.1145/3626772.3657837.

[27] G. Cheng, C. Jin, W. Ding, D. Xu, Y. Qu, Generating illustrative snippets for open data on the web, in: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, WSDM 2017, Cambridge, United Kingdom, February 6-10, 2017, ACM, 2017, pp. 151–159. URL: https://doi.org/10.1145/3018661.3018670. doi:10.1145/3018661.3018670.

[28] D. Liu, G. Cheng, Q. Liu, Y. Qu, Fast and practical snippet generation for RDF datasets, ACM Trans. Web 13 (2019) 19:1–19:38. URL: https://doi.org/10.1145/3365575. doi:10.1145/3365575.

[29] X. Wang, G. Cheng, T. Lin, J. Xu, J. Z. Pan, E. Kharlamov, Y. Qu, PCSG: pattern-coverage snippet generation for RDF datasets, in: The Semantic Web - ISWC 2021 - 20th International Semantic Web Conference, ISWC 2021, Virtual Event, October 24-28, 2021, Proceedings, volume 12922 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 3–20. URL: https://doi.org/10.1007/978-3-030-88361-4_1. doi:10.1007/978-3-030-88361-4_1.