Quantifying Ranker Coverage of Different Query Subspaces

Negar Arabzadeh
University of Waterloo
narabzad@uwaterloo.ca

Amin Bigdeli
Toronto Metropolitan University
abigdeli@torontomu.ca

Radin Hamidi Rad
Toronto Metropolitan University
radin@torontomu.ca

Ebrahim Bagheri
Toronto Metropolitan University
bagheri@torontomu.ca

ABSTRACT

The information retrieval community has observed significant performance improvements over various tasks due to the introduction of neural architectures. However, such improvements do not necessarily seem to have happened uniformly across a range of queries. As we will empirically show in this paper, the performance of neural rankers follow a long-tail distribution where there are many subsets of queries, which are not effectively satisfied by neural methods. Despite this observation, performance is often reported using standard retrieval metrics, such as MRR or nDCG, which capture average performance over all queries. As such, it is not clear whether reported improvements are due to incremental boost on a small subset of already well-performing queries or addressing queries that have been difficult to address by existing methods. In this paper, we propose the `Task Subspace Coverage` (`TaSC` /tAHsk/) metric, which systematically quantifies whether and to what extent improvements in retrieval effectiveness happen on similar or disparate query subspaces for different rankers. We show that the consideration of our proposed `TaSC` metric in conjunction with existing ranking metrics provides deeper insight into ranker performance and their contribution to overall advances.

CCS CONCEPTS

• **Information systems → Information retrieval**; **Query intent**; **Users and interactive retrieval**; **Collaborative search**; *Specialized information retrieval*; **Retrieval effectiveness**.

KEYWORDS

Evaluation, Retrieval Effectiveness, Information Retrieval Benchmarks

ACM Reference Format:
Negar Arabzadeh, Amin Bigdeli, Radin Hamidi Rad, and Ebrahim Bagheri. 2023. Quantifying Ranker Coverage of Different Query Subspaces. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3539618.3592045

SIGIR '23, July 23–27, 2023, Taipei, Taiwan
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9408-6/23/07…$15.00
https://doi.org/10.1145/3539618.3592045

1 INTRODUCTION

Neural rankers have shown increased effectiveness particularly on tasks such as ad hoc retrieval [21, 22], question-answering [10, 19], expert search [15–18] and so on. Despite their increased performance, neural rankers still face certain limitations. Most notably, researchers have found that while neural rankers have shown consistent performance improvement, such improvements are primarily due to their success with a specific subset of queries and not necessarily generalizable to a whole range of queries. The work by Arabzadeh et al. [1] was among the first to identify this issue by showing that a significant number of queries on the MS MARCO passage retrieval task remain completely unaddressed (reciprocal rank of zero) by neural rankers. One of the reasons for this could be the fact that most Information Retrieval (IR) tasks adopt a standard evaluation metric, such as MRR or NDCG. Such standard metrics are used to measure the success of any new method for that task; therefore, incentivizing researchers to focus on optimizing their methods for one or a specific set of metrics. Although standard metrics enable a fair comparison across different methods, they do not necessarily encourage researchers to work on the more difficult aspects of each task [2]. For instance, within the context of the MS MARCO passage retrieval task, existing neural methods seem to have focused on optimizing the performance of a certain subset of queries that would result in overall increased MRR@10 while overlooking another extremely difficult subset of queries (around 40% of the query set), which are non-trivial to address [1].

Given such non-uniform performance across different subsets of queries, researchers have advocated for the need to evaluate ranking methods specifically on difficult queries [7, 13, 20]. An improved ranker would be beneficial to the community if it not only shows improved average performance improvements over all queries in the standard dataset but also on the subsets of the queries that are known to be difficult. We find this idea to be quite strong as it will show how much breakthrough a method has been able to make on queries that have been traditionally harder for existing rankers to address. However, given the fact that the set of difficult queries may vary depending on the neural ranker, it may not be trivial to identify, and maintain a standard set of difficult queries that would be universally used to evaluate various neural rankers. This necessitates a standard quantitative approach for measuring the performance of ranking methods on difficult queries without having to rely on ranker-specific difficult queries, whose results may not be comparable to other rankers.

We advocate for the need to evaluate ranking methods from both (1) an overall perspective through standard metrics over a large test collection; as well as, (2) a more nuanced perspective focusing on

Table 1: The list of rankers used in our studies and their performance on the dev set of the MS MARCO passage ranking task.

Ranker	Citation	Released on	MRR@10	NDCG@10	Retrieval Strategy
RepBERT	[23]	June 2020	0.2968	0.3521	Uses fixed length contextualized embeddings to represent documents and queries and calculates the relevance of documents for queries using their inner product.
ANCE	[21]	October 2020	0.3304	0.3879	Distinguishes itself from other retrievers by using negative samples from an approximate nearest neighbor index and calculates the relevance of document-query pairs using the dot product of their learned dense representations.
SBERT	[19]	March 2021	0.3439	0.3905	Shown to be fast and scalable, and can efficiently search large collections of text using a siamese network architecture to encode the passages and the collection.
ColBERT (+Hybrid)	[11]	October 2020	0.3350 (0.3529)	0.3940 (0.4168)	Uses a cost-effective interaction step to model the similarity. This pruning-friendly mechanism reduces the cost of re-ranking documents and allows for end-to-end retrieval. We additionally consider the hybrid version of Colbert with a sparse retriever.
UniCOIL	[12]	June 2021	0.3509	0.4117	a sparse retriever which is a simple extension of COIL that produces representations for each document token that are then directly stored in an inverted index.
SPLADE	[8]	September 2021	0.3684	0.433	Provides highly sparse representations that could inherit from the desirable properties of bag-of-words models such as the exact matching of terms and the efficiency of inverted indexes.

less explored and more difficult queries. To this end, we propose the 'Task Subspace Coverage' (TaSC) metric that captures the extent to which a new ranker is able to address those subspaces that have not been effectively addressed by earlier rankers. Our proposed TaSC metric quantitatively shows how well a new method is able to satisfy query subsets that were deemed to be difficult for other rankers. Such information cannot be captured by *statistical significance tests*. When used in tandem with standard metrics, our proposed metric quantifies the performance of a given ranker on the query subspaces that are difficult for other rankers, while the standard metrics offer a more overall view of the method performance; hence together, they offer complementary perspectives.

2 EMPIRICAL EVIDENCE

Let us motivate the need for our proposed TaSC metric by exploring the MS MARCO passage retrieval task which [14] is designed for the sake of training neural models for ad hoc retrieval tasks. MS MARCO adopts MRR@10 as its standard evaluation metric. We adopt a set of seven different first-stage retrieval models, the details of which are included in Table 1. For the sake of clarity, we plot the sorted performance of each ranker in terms of MRR@10 for each individual query in Figure 1. It is shown that all the rankers ranging from dense retrievers to sparse learnt representations, suffer from long-tailed performance where there are over 35% of queries have an MRR@10 value of zero across all these rankers. Furthermore, while the performance of many of these rankers on the standard task metric, i.e., MRR@10, is quite competitive, this does not necessarily tell us how such performances are obtained and whether these methods impact the same or a different subset of queries. In Figure 2, we plot the pairwise difference of performance of four of these rankers on a per-query basis, which shows the extent to which they overlap with one another. A positive or a negative value in the Figure for each query shows that one of the rankers had a better performance on that particular query, and a value of zero shows that the two rankers performed exactly the same on that query. By contrasting the performance effectiveness values in Table 1 with Figure 2, one can see that while various methods have reported competitive performance, they do not necessarily show similar behavior over the queries. For instance, the standard MRR@10 metric shows that ColBERT (0.3350), ANCE (0.3304) and SBERT (0.3439) are competitive, but the in-depth analysis in Figure 2 shows that ColBERT and SBERT are addressing very similar query subspaces while ANCE is addressing a rather different subspace compared to these two rankers.

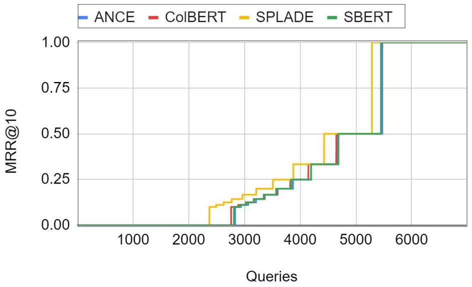

Figure 1: Various rankers' performance per-query.

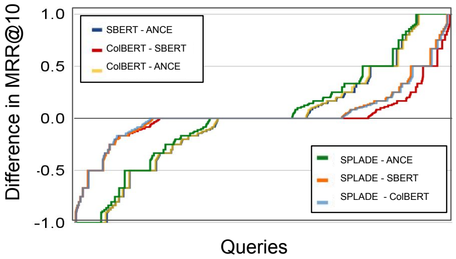

Figure 2: Pairwise comparison of rankers per-query.

Having an in-depth understanding of the overlap between the subspaces addressed by each ranker would be useful for understanding how the ranker is making breakthroughs on queries that were harder for others. This would be beneficial for building stronger rankers through ensemble methods [4–6] or query routing [3, 9].

The objective of our work is to propose the Task Subspace Coverage (TaSC) metric, which would systematically capture to what extent a given ranker is addressing queries that are harder for other rankers to address. We advocate that the TaSC metric could accompany any task's official metric in order to provide a fuller picture of the overall method performance as well as its impact on query subspaces that are deemed more difficult for other rankers.

3 THE PROPOSED TaSC METRIC

The objective of TaSC metric is to supplement existing standard metrics by allowing to not only track overall performance improvements shown by recent methods, but also identify whether the new method is able to explore and effectively address those subspaces that are difficult for existing state of the art methods. The idea behind our proposed metric is quite intuitive and directly speaks to the need for newer methods to address subspaces that are less explored by earlier methods. TaSC provides the means to quantitatively measure to what extent a new method is able to satisfy those queries that were not dealt with satisfactorily by other methods. Let us provide a formal treatment of TaSC as follows: Given a query q, a collection of items C and an information retrieval method R, we define $R(q, C) = D_q$ where $D_q = [d_1, d_2, ...d_k]$ is the retrieved ranked list of k items $d_i \in C$ by R in order to satisfy the information need behind query q. R could be evaluated on a per query basis through an evaluation metric μ where $\mu\{q, d_q|J_q\} \longrightarrow [0, 1]$ is an evaluation metric that maps the performance of retrieval method R onto a scalar value based on the set of relevant judged items J_q. The overall performance of retrieval method R, $Eval(R)$ is obtained by assessing R on a set of queries $Q = \{q_1, q_2, ...q_n\}$ through the aggregated quantified performance over all the queries in the set:

$$Eval(R) = \frac{1}{|Q|}\sum_{q \in Q} \mu(q, d_q|J_q) \tag{1}$$

Let $\mathcal{R}$ be a sequenced list of rankers $[R_{t_1}, R_{t_2}, \ldots, R_{t_n}]$ such that retrieval method R_{t_i} is proposed prior to retrieval method $R_{t_{i+1}}$. The objective of TaSC is to quantify how different a retrieval method, such as R, is compared to $\mathcal{R}$ which is the set of rankers that have been proposed prior to R. A highly similar behaviour of R to $\mathcal{R}$ shows that the new method addresses similar query subspaces to those explored by prior retrieval methods in $\mathcal{R}$. Otherwise, R is exploring novel subspaces. We define TaSC for ranker R based on a list of rankers $\mathcal{R} = [R_{t_1}, R_{t_2}, \ldots, R_{t_n}]$ and on the query set Q as:

$$\texttt{TaSC}_{agg}(R_{t_i}, Q) =$$
$$\frac{1}{|Q|}\sum_{q \in Q}(1 - agg\{\mu(q, R_{t_j}|J_q)|R_{t_j} \in R, j < i\}) \times \mu(q, R_{t_i}|J_q) \tag{2}$$

where agg is an aggregation function such as average or maximum. TaSC metric computes the degree of difficulty of queries in the query space based on the performance of earlier retrieval methods on those queries. TaSC exploits the degree of difficulty of each query based on the performance of past retrieval methods to discount the performance observed by the new method i.e., easier queries for past methods receive a lower degree of importance in TaSC while more difficult queries based on methods in $\mathcal{R}$ receive a higher weight and importance. As an example, a query q that received a maximum performance metric value of zero by existing methods, indicating that it was an extremely difficult query, would receive a weight of 1 in TaSC while a query with a maximum metric value of 1 from methods in $\mathcal{R}$ would not be taken into consideration by TaSC, as this query has already been fully satisfied by earlier methods.

In summary, the proposed TaSC metric provides a weighted treatment of the standard evaluation metric over the whole query set where the weights are determined based on the difficulty of each query for existing state-of-the-art baselines. On this basis, TaSC is able to show whether a new method is showing performance improvements due to exploring similar subspaces to existing methods or because it is exploring newer and less explored subspaces.

4 OBSERVATIONS AND DISCUSSION

Here, we are interested in showing the impact of TaSC in depicting the performance of the rankers introduced in Table 1. To do so, we order the rankers based on their release date and compute the TaSC metric for each of these rankers based on the performance of the rankers that were released before them. For instance, RepBERT is the earliest model included in our set of rankers and since there are no prior methods in our set of methods in Table 1, we consider BM25 as the baseline method for RepBERT. For RepBERT $\mathcal{R} = [BM25]$ and similarly, for ANCE, which was released next in Table 1, $\mathcal{R} = [BM25, RepBERT]$. For the latest ranker in our experiment, i.e., SPLADE, all the rankers in Table 1 as well as BM25 were included in $\mathcal{R}$. Figure 3 shows the TaCS metric values for different retrieval methods. We note that retrieval methods are ordered chronologically on the x-axis based on their release date.

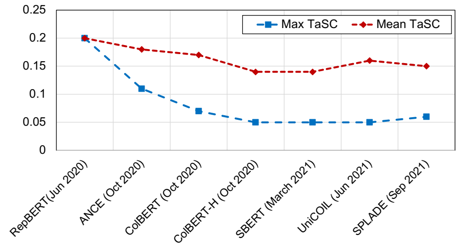

Figure 3: The value of our proposed TaSC metric over various retrieval methods released over time.

Figure 3 allows us to make several observations: **(1)** The TaSC metric for the RepBERT method is the highest among all other methods. This is primarily because the performance of RepBERT, which is a neural ranker, is compared with BM25, which is a sparse retriever. The high TaSC value indicates that RepBERT has not only been able to improve performance on MRR@10 from 0.1874 (BM25) to 0.2968, but also this increase in performance is not due to improvement of only easier queries for BM25. It in fact happened because RepBERT was able to address query subsets that were not accessible to BM25. **(2)** After RepBERT, the general trend of TaSC values on subsequent neural methods up to and including SBERT is decreasing, which indicates that these methods have been increasingly focused on a smaller subset of queries that are considered to be easier for earlier methods. The small TaSC values on SBERT and ColBERT are indications that while the MRR@10 values of these methods are quite strong (0.3350 and 0.3439), these improvements are a result of impact on queries that have already been addressed well by earlier methods. **(3)** The contrast between the TaSC values computed based on the maximum and average aggregation functions point to the fact that in many cases improvements observed by neural rankers is due to improvements on queries that have already been partially addressed by earlier methods. **(4)** Finally, we note an increase in TaSC values especially by SPLADE, which deviates from the neural ranker paradigm be learning sparse representations to perform retrieval. This shift shows both performance improvement over the best previous baseline (0.3509 vs 0.3684) on the standard MRR@10 metric and increased TaSC value indicating that the improved performance is a result of exploring query subspaces that have not been addressed by earlier rankers.

We further illustrated the ternary relationship between the release date of each ranker, the ranker's retrieval effectiveness MRR@10, and our proposed TaSC metric in Figure 4(a) and (b) in which the X-axis presents the release date of the method, the Y-axis presents the TaSC metric and the size of the bubble for each ranker presents its MRR@10 value. The consideration of TaSC metric allows us to present several findings: **(1)** our first finding relates to the relationship between the standard evaluation metric and TaSC. While many

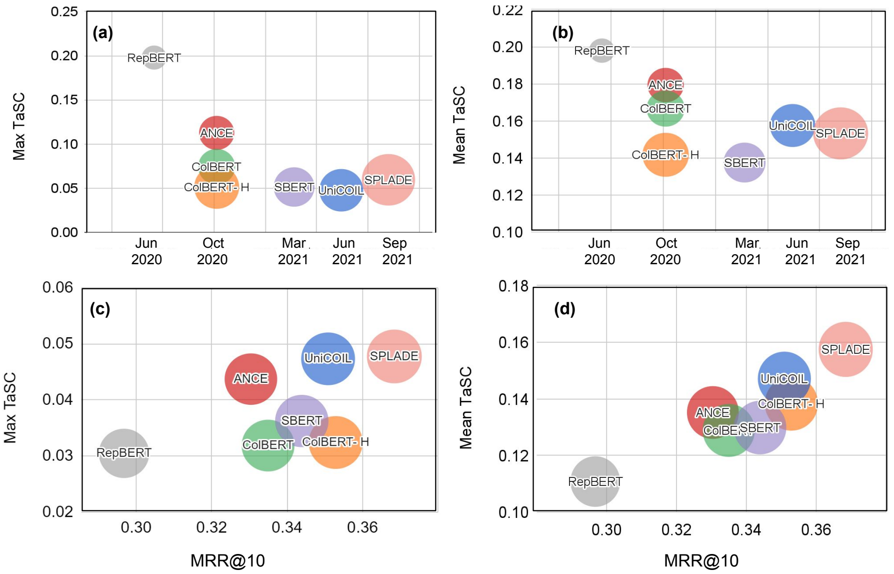

Figure 4: (a) and (b) : The ternary relation between ranker release date (x-axis), TaCS values (y-axis), and their MRR@10 values (bubble size). (c) and (d) : TaSC metric values when rankers are compared to all other rankers regardless of date of release.

of the methods have comparable performance on the standard metric (MRR@10), their TaSC values are notably different. For example, when comparing ANCE, ColBERT and ColBERT-Hybrid, and based on Table 1, their MRR@10 values are comparable at 0.3304, 0.3350, and 0.3529, respectively, with ANCE showing the lowest performance on MRR@10. However, ANCE has the highest TaSC value, which is an indication that it has been able to obtain 0.3304 on MRR@10 by exploring query subspaces that are not necessarily overlapping with those addressed by RepBERT. In contrast, while ColBERT-Hybrid has the highest MRR@10 among the three methods (0.3529), it shows the lowest TaSC value. This means that the queries improved by ColBERT-Hybrid are quite overlapping with those addressed by RepBERT. **(2)** When exploring the Max TaSC in Figure 4 (a), we observe that SPLADE has a noticeable MRR@10 value (size of bubble) and at the same time exhibits an increased TaSC value. The increased TaSC value on Max TaSC is an indication that the method is performing increasingly well on the difficult queries for the earlier baselines. **(3)** In the context of Mean TaSC in Figure 4 (b), the notable improvement on TaSC by SPLADE and UniCOIL shows that these methods have been able to substantially improve the performance of those queries that were only satisfied by other methods to a limited extent. Higher Mean TaSC metric values indicate that SPLADE and UniCOIL are able to boost the performance of queries that were only partially addressed by earlier methods. Overall, one could conclude that a method such as SPLADE has shown its impressive retrieval effectiveness measured by MRR@10 by improving a subset of extremely difficult queries for the baselines (shown by Max TaSC), as well as also improving the performance of a subset of queries that were only partially addressed by the baseline methods (shown by Mean TaSC).

While we have advocated for the importance of our proposed TaSC metric to track the impact of newer retrieval methods on exploring more difficult query subspaces for existing methods, it is important to highlight that our proposed metric can also be used to quantitatively measure to what extent any given method covers query subspaces that are more difficult for other methods regardless of when those methods were released. Given a retrieval method R, one can define $\mathcal{R}$ to include any set of other baselines regardless of when they were released. In Figure 4 (c) and (d) the TaSC metric is computed for each method by comparing it to the rest of the retrieval methods of Table 1. We make several observations based on the TaSC metric behavior in this Figure: **(1)** We notice that while RepBERT had quite a high TaSC value when compared to BM25, it shows lower TaSC value when compared to other rankers i.e., the other methods successfully cover most of the queries that are satisfied by RepBERT and when factoring the MRR@10 value for this method (x-axis), one may conclude that RepBERT is not an ideal choice. **(2)** From among the methods that have comparable MRR@10 performance, namely UniCOIL, ColBERT-H and SBERT, UniCOIL shows the highest TaSC value, which indicates that it covers a larger number of queries that are deemed to be difficult for the other methods while the lower TaSC value of SBERT and ColBERT-H shows that they are covering similar query subsets. **(3)** From among all the rankers, SPLADE and UniCOIL which are both based on learnt sparse representations, exhibit the highest MRR@10 values as well as the largest TaSC values. This indicates that these two methods are those that show good performance overall but also on query subsets that are difficult for other rankers. As such, they could be ideal candidates to be used for building hybrid rankers.

5 CONCLUDING REMARKS

We proposed the Task Subspace Coverage (TaSC) metric, which captures to what extent retrieval methods cover similar or disparate query subspaces. We showed that reporting standard retrieval effectiveness metrics, such as MRR@10, along with our proposed TaSC metric can lead to additional insights that would not otherwise be accessible. TaSC allows for the understanding of whether any given retrieval method obtains better retrieval effectiveness through further improving already addressed queries by existing baselines but with greater performance, or that such improved effectiveness is due to the exploration and satisfaction of queries from less or totally unexplored query subspaces of other methods.

REFERENCES

[1] Negar Arabzadeh, Bhaskar Mitra, and Ebrahim Bagheri. 2021. MS MARCO Chameleons: Challenging the MS MARCO Leaderboard with Extremely Obstinate Queries. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management.* 4426–4435.

[2] Negar Arabzadeh, Alexandra Vtyurina, Xinyi Yan, and Charles L. A. Clarke. 2021. Shallow pooling for sparse labels. *CoRR* abs/2109.00062 (2021). arXiv:2109.00062 https://arxiv.org/abs/2109.00062

[3] Negar Arabzadeh, Xinyi Yan, and Charles LA Clarke. 2021. Predicting efficiency/effectiveness trade-offs for dense vs. sparse retrieval strategy selection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management.* 2862–2866.

[4] Steven M Beitzel, Eric C Jensen, Abdur Chowdhury, David Grossman, Ophir Frieder, and Nazli Goharian. 2004. Fusion of effective retrieval strategies in the same information retrieval system. *Journal of the American Society for Information Science and Technology* 55, 10 (2004), 859–868.

[5] Steven M Beitzel, Eric C Jensen, Abdur Chowdhury, David A Grossman, Nazli Goharian, and Ophir Frieder. 2003. Recent Results on Fusion of Effective Retrieval Strategies in the Same Information Retrieval System.. In *Distributed Multimedia Information Retrieval.* Springer, 101–111.

[6] Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval.* 758–759.

[7] Faezeh Ensan and Weichang Du. 2019. Ad hoc retrieval via entity linking and semantic similarity. *Knowledge and Information Systems* 58 (2019), 551–583.

[8] Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE v2: Sparse lexical and expansion model for information retrieval. *arXiv preprint arXiv:2109.10086* (2021).

[9] Hai Jin, Xiaomin Ning, Hanhua Chen, and Zuoning Yin. 2006. Efficient query routing for information retrieval in semantic overlays. In *Proceedings of the 2006 ACM symposium on Applied computing.* 1669–1673.

[10] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906* (2020).

[11] Omar Khattab and Matei Zaharia. 2020. 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.* 39–48.

[12] Jimmy Lin and Xueguang Ma. 2021. A few brief notes on deepimpact, coil, and a conceptual framework for information retrieval techniques. *arXiv preprint arXiv:2106.14807* (2021).

[13] Sean MacAvaney, Andrew Yates, Arman Cohan, Luca Soldaini, Kai Hui, Nazli Goharian, and Ophir Frieder. 2019. Overcoming low-utility facets for complex answer retrieval. *Information Retrieval Journal* 22 (2019), 395–418.

[14] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *choice* 2640 (2016), 660.

[15] Radin Hamidi Rad, Ebrahim Bagheri, Mehdi Kargar, Divesh Srivastava, and Jaroslaw Szlichta. 2021. Retrieving Skill-Based Teams from Collaboration Networks. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 2015–2019. https://doi.org/10.1145/3404835.3463105

[16] Radin Hamidi Rad, Hossein Fani, Ebrahim Bagheri, Mehdi Kargar, Divesh Srivastava, and Jaroslaw Szlichta. 2023. A Variational Neural Architecture for Skill-Based Team Formation. *ACM Trans. Inf. Syst.* (April 2023). https://doi.org/10.1145/3589762

[17] Radin Hamidi Rad, Hossein Fani, Mehdi Kargar, Jaroslaw Szlichta, and Ebrahim Bagheri. 2020. Learning to Form Skill-based Teams of Experts. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux (Eds.). ACM, 2049–2052. https://doi.org/10.1145/3340531.3412140

[18] Radin Hamidi Rad, Shirin Seyedsalehi, Mehdi Kargar, Morteza Zihayat, and Ebrahim Bagheri. 2022. A Neural Approach to Forming Coherent Teams in Collaboration Networks. In *Proceedings of the 25th International Conference on Extending Database Technology, EDBT 2022, Edinburgh, UK, March 29 - April 1, 2022*, Julia Stoyanovich, Jens Teubner, Paolo Guagliardo, Milos Nikolic, Andreas Pieris, Jan Mühlig, Fatma Özcan, Sebastian Schelter, H. V. Jagadish, and Meihui Zhang (Eds.). OpenProceedings.org, 2:440–2:444. https://doi.org/10.48786/edbt.2022.37

[19] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[20] Xuanhui Wang, Hui Fang, and ChengXiang Zhai. 2007. Improve retrieval accuracy for difficult queries using negative feedback. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management.* 991–994.

[21] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808* (2020).

[22] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. Learning to retrieve: How to train a dense retrieval model effectively and efficiently. *arXiv preprint arXiv:2010.10469* (2020).

[23] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. Repbert: Contextualized text embeddings for first-stage retrieval. *arXiv preprint arXiv:2006.15498* (2020).