Evaluating Fusion Functions and Normalization Strategies for Multi-Model Information Retrieval - Group 4

Danae Natalie Savvidi
Delft University of Technology
Delft, Netherlands

Henry Page
Delft University of Technology
Delft, Netherlands

Bernadett Bakos
Delft University of Technology
Delft, Netherlands

Weicheng Hu
Delft University of Technology
Delft, Netherlands

Abstract

Hybrid information retrieval (IR) aims to combine the precision of lexical methods with the semantic understanding of neural models. While effective in principle, integrating outputs from multiple retrievers poses a challenge: how best to fuse the diverse scoring signals. This paper investigates the performance of fusion functions in multi-model hybrid IR, focusing on rank-based Reciprocal Rank Fusion (RRF), and score-based Convex Combination (CC) techniques. We also assess how normalization affects CC performance. Experiments on four BEIR datasets show that normalization can yield consistent gains, with CC generally outperforming RRF.

CCS Concepts

• **Information systems → Rank aggregation**; *Combination, fusion and federated search*; Learning to rank.

Keywords

Hybrid Retrieval, Fusion Functions, Normalization in Convex Combination, Lexical and Semantic Search

1 Introduction

There are three main approaches to information retrieval: sparse, dense, and hybrid. Sparse retrieval methods, such as BM25 [12] and TF-IDF [20], represent documents as sparse vectors, relying on exact term matching. While these methods are computationally efficient, they often struggle with vocabulary mismatch, which limits their effectiveness in capturing semantic meaning [1]. In contrast, dense retrieval techniques, such as those using BERT and Llama, leverage dense embeddings to capture deeper semantic relationships. However, despite their improved retrieval performance, these methods demand significantly higher computational resources.

To balance the strengths and weaknesses of sparse and dense retrieval, hybrid retrieval systems integrate both lexical and semantic signals. However, an essential challenge in hybrid retrieval is the selection of an appropriate fusion function to combine the results from different ranking models. Two common fusion methods include Reciprocal Rank Fusion (RRF) [4] and Convex Combination (CC). Authors in a recent study [3] claim that convex combination is sensitive to its parameter α and the choice of score normalization. This would make RRF - a non-parametric fusion function - more suitable. Another study [2], re-examined those findings and found that the choice of score normalization does not significantly affect ranking performance. However, the authors of that paper leave open the question of how these conclusions might scale when three or more retrieval models are involved.

This study builds upon those findings and aims to address the following research questions:

- **RQ 1:** How do different fusion functions perform when combining the results of three or more retrieval models?
- **RQ 2:** How does the choice of normalization affect the performance of convex combination in model fusion with three or more models, and how does it compare to RRF?

This paper is structured as follows: Chapter 2 provides a summary of related literature on fusion functions and score normalization. In Chapter 3, we describe the methodology used in our experiments. Chapter 4 explains the experiments we conducted, the corresponding results, and an analysis on the limitations and implications of our findings. Finally, Chapter 5 summarizes our conclusions.

2 Related Work

2.1 Fusion Functions

2.1.1 Score-Based. The integration of sparse and dense retrieval techniques into hybrid retrieval systems has been an emerging topic. Numerous studies have explored different strategies for combining the outputs of sparse (lexical) and dense (semantic) retrieval models across diverse domains.

A popular approach involves combining BM25 as the lexical component with a variant of BERT as the semantic model [8, 13, 15, 16, 23]. Most of these works fuse the results of the two models using the vector dot-product as a similarity measure [8, 15, 16, 23], while Kuzi et al. [13] utilize the relevance model RM3 [11] for merging results. Across all these studies, hybrid models consistently outperform individual retrieval methods, underscoring the complementary strengths of lexical and semantic approaches.

2.1.2 Rank-Based. RRF, a rank-based approach, has been widely adopted in the literature for various applications. For instance, Nunzio et al. [19] explored several configurations of three models,

DSAIT4050-Q3-25, TU Delft

using both RRF and a hybrid of RRF and RM3. Their findings show that the RRF-RM3 combination yielded the best results, highlighting the potential of fusing more than two models.

RRF has also been used in large-scale ensemble systems [24]. Cormack et al. [5] showed that RRF outperforms both Condorcet [18] and individual best fusing when aggregating results from 30 different Wumpus Search settings. However, these studies were limited to RRF and did not investigate the effectiveness of linear fusion methods.

2.2 Normalization Methods

The impact of normalization on fusion strategies has been explored in prior work. Song et al. [21] conducted experiments evaluating several normalization techniques, namely the fitting method, Zero-one [14], Sum, and ZMUV [17] in combination with the CombSum and CombMNZ fusion methods [7]. Their results demonstrate that the choice of normalization strategy, when paired with different fusion techniques, can influence performance. However, their study was limited to a subset of possible normalization methods and fusion functions.

The above mentioned studies focus mainly on in-domain scenarios. In contrast, Chen et al. [3] examine hybrid retrieval performance in both in-domain and out-of-domain contexts. Their findings indicate that semantic models degrade significantly when applied to out-of-domain datasets, while lexical models maintain robust performance. Their proposed hybrid model (of two retrievers) outperforms both individual components. In their experiments, they compare the performance of RRF and linear combination (CC) with min-max scaling and find that RRF outperforms linear combination across all settings.

Bruch et al. [2] conduct an analysis of fusion functions and the role of normalization on out-of-domain retrieval tasks when fusing two retrievers. Contrary to the findings of Chen et al. [3], they report that CC outperformed RRF in both in-domain and out-of-domain settings. Moreover, they argue that the choice of normalization method has a negligible effect on overall performance, as appropriate tuning of hyperparameters can offset its impact.

Given the conflicting conclusions of these studies, our work aims to further investigate hybrid retrieval by combining more than two retrieval models using different fusion techniques. We also examine whether the choice of normalization strategy significantly influences retrieval performance in such a setting.

3 Method

We provide an overview of our methodology in Figure 1. The query is first processed using BM25 to retrieve the top 1000 documents. These are then re-ranked for semantic relevance using DistilBERT and Contriever. Then, the three scores are normalized and fused, but normalization of scores happens only when using normalized CC. As we build upon the research of [2], we adopt the same datasets, fusion functions, normalization techniques, and metrics to ensure that our results are directly comparable to those previously obtained.

3.1 Retrievers

Lexical retrieval: For lexical retrieval, we use BM25 [12], a well-established probabilistic ranking function that scores documents based on term frequency and inverse document frequency. It is an efficient technique and gives strong performance on keyword-based retrieval tasks. We set its hyperparameters to ($k1 = 0.9$ and $b = 0.4$) following [2, 3], to retrieve the top 1000 candidates.

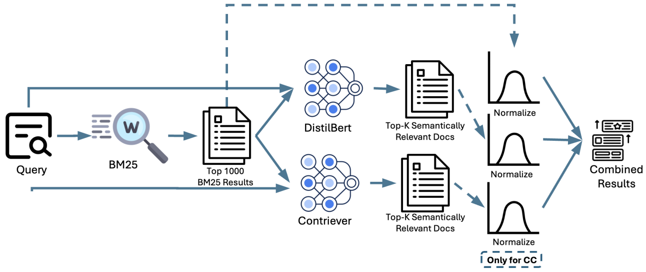

Figure 1: Overview of the methodology.

Semantic retrieval: For semantic retrieval, we leverage a modified version of BERT [6] called DistilBERT [9], as it is available in the PyTerrier framework used for implementation. BERT encodes queries and documents into dense vector representations, enabling retrieval based on semantic similarity. To answer our research questions, we need three models. One option is to use two lexical models and one semantic model; however, since BM25 is already the best-performing lexical model, adding an additional lexical model would not improve retrieval performance significantly. Instead, we opt to add another semantic model: Facebook's Contriever [10].

3.2 Fusion Functions & Normalization Techniques

We experiment with various normalization techniques used in CC and compare the results to those obtained using RRF.

Convex Combination (CC). Convex Combination (Eq. 1) can be seen as a weighted average of the lexical and semantic scores.

$$
\begin{aligned}
f_{Convex}(q,d) = \alpha f_{BM25}(q,d) + \beta f_{Contriever}(q,d) \\
+ (1 - \alpha - \beta) f_{DistilBert}(q,d)
\end{aligned}
\tag{1}
$$

Subject to $\alpha, \beta \geq 0, \quad \alpha + \beta \leq 1$, where α, β are weight parameters that determine the relative importance of each method. Note, due to different score distributions, we need to perform normalization for all models.

We explore the following normalization techniques for the hybrid CC method (modified from [2]):

- $\phi_{\mathbf{mm}}$: Min-Max normalization scales scores to a range [0, 1] based on observed minimum and maximum values.
- $\phi_{\mathbf{tmm}}$: Theoretical Min-Max normalization uses predefined minimum and maximum values.
- $\phi_{\mathbf{z}}$: Z-score normalization centers scores around 0 with a standard deviation of 1 by subtracting the mean and dividing by the standard deviation.
- $\phi_{\mathbf{mm\text{-}Lex}}$: Min-Max scaling applied only to lexical scores, leaving semantic scores unnormalized.
- $\phi_{\mathbf{tmm\text{-}Lex}}$: Theoretical Min-Max scaling applied to lexical scores, with semantic scores unnormalized.

- $\phi_{\mathbf{z\text{-}Lex}}$: Z-score normalization applied only to lexical scores, leaving semantic scores unnormalized.
- $\mathbb{I}$: Identity transformation leaves both lexical and semantic scores unnormalized.

Reciprocal Rank Fusion (RRF). RRF ensures that even documents ranked moderately well across multiple models receive a higher final score than those ranked highly in only one model. The RRF score for a document d and a query q is computed as follows:

$$\mathrm{RRF}(q,d) = \sum_{i=1}^{R=3} \frac{1}{\eta_i + \mathrm{rank}_i(d,q)}$$

where $\mathrm{rank}_i(q,d)$ is the rank of document d in the i-th retrieval model, and η_i is a constant used to control the influence of document ranks. As done in [2, 3], we take a non-parametric view of RRF and set η_i to its default value 60. This prevents very low-ranked documents from contributing excessively to the final score.

3.3 Evaluation

3.3.1 Datasets. We evaluate our methods on four publicly available datasets from the BeIR collection [22]: FiQA (Finance Q&A), SciFact (scientific claim verification), Quora (duplicate detection), and NFCorpus (medical). Although testing on the entire BeIR collection would be ideal, computational constraints limit us to a few representative datasets. We select our datasets to ensure they cover a broad range of topics and query types. A more detailed description of each dataset can be found in [22].

3.3.2 Metrics. To evaluate the ranking quality and relevance that different fusion functions produce, we use two metrics, Normalized Discounted Cumulative Gain (NDCG) and Recall. NDCG measures the ranking quality by considering the relevance of retrieved documents while giving higher importance to documents appearing earlier in the ranking. Recall measures the ability of a ranking system to retrieve relevant documents.

4 Experiments

This chapter begins our experimental setup, including pipeline logistics, fusion details, and baselines. Following this, we analyze the results for each experiment and highlight their limitations as well as their implications.

4.1 Experimental Setup

Pipeline Setup. We first index the training sets, and then use these to retrieve documents from the test set for each model. We created 3 pipelines: $BM25$, $BM25 >> DistilBert$ and $BM25 >> FacebookContriever$. BM25 will retrieve the Top 1000 result documents for a given query. For the Quora Dataset, due to its scale and our hardware constraints, the cutoff is set to 50, and R@10, NDCG@10 is used instead of R@100, NDCG@100. When BM25 finishes the initial retrieval, the neural models act as a re-ranker separately.

Model Fusion. We combine retrieval results from the three models using different normalization tactics for convex combination, with weights 0.2 for pure $BM25$, 0.4 for $BM25 >> DistilBert$, and 0.4 for $BM25 >> FacebookContriever$. We chose these hyperparameters based on the findings of Bruch et al. [2], where they concluded that assigning a weight of 0.8 to the semantic model performs generally well across these datasets. Since our approach incorporates two semantic models, we opted to distribute this weight equally between them, giving both models 0.4.

Baselines. We selected the Convex Combination with Identity transformation ($\mathbb{I}$), along with RRF as our baseline functions. The identity transformation is included to examine the role of normalization in fusion functions under controlled conditions, helping us address RQ1. RRF, a simple yet widely used fusion method shown to be effective in our reference study [2], serves as a strong comparative baseline and supports our investigation of RQ2.

4.2 Experiment Results

The results of the experiments can be seen in Table 1. To analyze the results in better detail, we performed a statistical test of significance.

4.2.1 RQ1: Fusion Function Performance in Multi-Model Retrieval. For this experiment, we evaluated the performance of the convex combination with different normalization strategies. Among all the methods tested, ϕ_{tmm} and ϕ_z consistently achieved the best NDCG@100 and Recall@100 scores across most datasets. This relatively stable performance despite the shifts in data distribution is consistent with findings from prior work [2], which suggested that ϕ_{tmm} is the most robust normalization method out of these six.

The study attributes this robustness to the fact that ϕ_{tmm} replaces the minimum score in the retrieved set with a fixed, feasible minimum value, instead of relying on the dataset-specific minimum score. This reduces the influence of data-dependent statistics, which might otherwise cause performance to fluctuate depending on the retrieval set.

In the results, however, we also see ϕ_z performing well. While we could not find literature explaining this phenomenon, we attempt to provide an explanation based on how z-score normalization works. In hybrid IR settings, especially when combining models with inherently different scoring scales, raw scores can be skewed or have very different variances. Min-max normalization compresses all scores into a $[0, 1]$ range, which may flatten meaningful differences if one model has a narrower range. In contrast, z-score normalization maintains the relative spacing between scores by scaling based on variance, not fixed bounds. This helps maintain rank-sensitive differences and avoids disproportionately suppressing signals from any single model. It is particularly effective when fusing models with asymmetric score distributions—a factor that becomes more impactful in multi-model settings.

Notably, the lexical-only versions of the convex combination (i.e., ϕ_{mm-Lex}, $\phi_{tmm-Lex}$, ϕ_{z-Lex}) consistently performed worse than their non-Lex counterparts. This is likely due to the fact that normalizing only the lexical scores, while leaving the semantic scores unnormalized, created an imbalance in the fusion process.

4.2.2 RQ2: Normalization Effects on Convex Combination. In our experiments, we found that convex combination remains a reliable fusion method, consistent with previous work [2]. The previous study showed that while normalization is important for consistency, the choice of normalization doesn't drastically affect performance

	Recall				NDCG			
Fusion	**NFCorpus**	**SciFact**	**FiQA**	**Quora**	**NFCorpus**	**SciFact**	**FiQA**	**Quora**
$\mathbb{I}$	0.254	0.937	0.633	0.918	0.288	0.719	0.386	0.851
ϕ_{mm}	**0.263**	0.942	0.653‡	0.924	0.295‡	0.718	**0.403***‡	0.862‡
ϕ_{tmm}	0.262	**0.952**	**0.659***‡	**0.925**	**0.296**‡	0.730	0.396*	**0.863**‡
ϕ_z	0.262	**0.952**	0.655‡	**0.925**	**0.296**‡	**0.733**	0.401*‡	**0.863**‡
ϕ_{mm-Lex}	0.253*	0.893*‡	0.606*‡	0.909	0.278*‡	0.660*‡	0.371‡	0.840‡
$\phi_{tmm-Lex}$	0.253*	0.893*‡	0.606*‡	0.909	0.278*‡	0.660*‡	0.371‡	0.840‡
ϕ_{z-Lex}	0.255	0.913*	0.621*‡	0.912	0.283*‡	0.695*‡	0.384	0.844‡
RRF (60,60)	0.262	0.948	0.643	0.922	0.292	0.730	0.380	0.853

Table 1: Mean NDCG@100 and Recall@100 (@10 for Quora) on the test split of four zero-shot datasets for hybrid retrieval using convex combination (CC) ($\alpha = 0.2, \beta = 0.4$) and RRF ($\eta_{LEX}, \eta_{SEM}$). The symbols ‡ and * indicate statistical significance (p-value < 0.01) with respect to unnormalized CC and baseline RRF (60, 60), respectively, according to a paired two-tailed t-test.

as long as the transformation is reasonable. However, with three models, we observed that normalization techniques like ϕ_{tmm} and ϕ_z provided small but consistent improvements. This suggests that although rank-equivalence still holds in theory, the practical differences introduced by each normalization method can become slightly more noticeable as the number of fused models increases.

Our statistical analysis supports the hypothesis that normalized CC generally outperforms its unnormalized counterpart. Furthermore, unlike prior findings, where RRF performed well at times [3], our results show that while RRF remained competitive and achieved better performance than convex combinations with lexical-only normalization, it was still consistently outperformed by normalized convex combinations, particularly ϕ_{tmm} and ϕ_z. It is important to note, however, that while the improvements over lexical-only CC were statistically significant, the differences between RRF and the fully normalized CC variants were often not.

4.3 Limitations & Implications

4.3.1 Limitations of findings. The study by Bruch et al. [2] suggests that the performance of the normalization function largely depends on the hyperparameters of the fusion function. However, due to computational limitations, we have considered only one parameter setting for the convex combination. Therefore, for further analysis on the effect of normalization, we suggest performing experiments with different settings for the weight α and β across the normalization strategies.

Given the constraints on time and resources, our experiments were limited to smaller, out-of-domain datasets. For future work, we recommend extending the evaluation to larger datasets and incorporating in-domain scenarios.

In our study, we explored the fusion of three different retrieval models: two semantic and one lexical. Future extensions of this work could involve experimenting with alternative model configurations, such as using one semantic and two lexical models, or incorporating more than three models to assess whether increasing the number of fused models impacts performance.

4.3.2 Implications of findings. Prior work often focused on combining two retrieval models, typically one lexical and one neural. Our results highlight that once you add a third model, especially another neural retriever, score normalization becomes more pivotal. This new insight adds to the literature by showing that even small differences in normalization strategies (such as Z-score vs. min–max) can lead to fairly consistent improvements in multi-model settings.

As more research groups incorporate multiple specialized retrievers, for example, domain-specific LLM or specialized medical IR systems, we provide evidence that normalizing all model outputs, instead of just lexical, generally yields better performance. This can be directly adopted by researchers and engineers looking to integrate new dense or sparse retrievers into existing pipelines.

On a broader level, our findings have the potential to influence a wide range of industries where information retrieval systems are crucial. As hybrid retrieval models gain traction, the importance of the fusion function and its normalization grows significantly. In sectors such as healthcare, law, and finance, where precision, reliability, and timely results are essential — advancements in information retrieval can lead to substantial improvements, ultimately driving more accurate, efficient, and impactful decision-making.

5 Conclusions

We investigated hybrid retrieval systems combining three models to evaluate fusion approaches in multi-model settings. Our key findings demonstrate that: (1) convex combination remains effective for three-model fusion when using proper normalization, with theoretical min-max and z-score performing best; and (2) comprehensive normalization of all model outputs becomes increasingly important as more models are combined, outperforming lexical-only normalization approaches.

These results advance hybrid retrieval research by showing that careful score normalization enables effective fusion of multiple retrieval models, while maintaining the benefits of both lexical and semantic approaches. The work particularly benefits scenarios requiring the integration of specialized retrievers into existing systems.

For future work, we identify three promising directions: (1) extension to larger model ensembles, (2) investigation of adaptive normalization techniques, and (3) application to domain-specific retrieval tasks where multiple specialized models must be combined.

6 Self-Assessment of Contributions & Student Numbers

Bernadett (5515890): Related Work and Limitations section.

Danae (5457769): Abstract, Introduction, Method sections and most of Experiment section. Run indexing for experiments.

Henry (5472636): Implement normalization techniques, Run experiments with RRF baseline, conclusion.

Weicheng (5225590): Experimental setup, procedure and implications. Setup repo template and implement fusion functions. Run experiments with identity baseline.

Link to Repository. https://github.com/danaesav/ir-project.

AI Disclaimer. We have used AI to rewrite some parts of our report for better phrasing. This included inputting our ideas and asking it to rewrite in more formal language.

References

[1] 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.

[2] Sebastian Bruch, Siyu Gai, and Amir Ingber. 2023. An analysis of fusion functions for hybrid retrieval. *ACM Transactions on Information Systems* 42, 1 (2023), 1–35.

[3] Tao Chen, Mingyang Zhang, Jing Lu, Michael Bendersky, and Marc Najork. 2022. *Out-of-Domain Semantics to the Rescue! Zero-Shot Hybrid Retrieval Models.* Springer International Publishing, 95–110. doi:10.1007/978-3-030-99736-6_7

[4] Gordon V. Cormack, Charles L A 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* (Boston, MA, USA) *(SIGIR '09).* Association for Computing Machinery, New York, NY, USA, 758–759. doi:10.1145/1571941.1572114

[5] Gordon V. Cormack, Charles L A 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* (Boston, MA, USA) *(SIGIR '09).* Association for Computing Machinery, New York, NY, USA, 758–759. doi:10.1145/1571941.1572114

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL] https://arxiv.org/abs/1810.04805

[7] E. A. Fox and J. Shaw. 1994. Combination of multiple searches.. In *The Second Text REtrieval Conference (TREC-2)* (Gaitherburg, MD, USA). 243–252.

[8] Luyu Gao, Zhuyun Dai, Tongfei Chen, Zhen Fan, Benjamin Van Durme, and Jamie Callan. 2021. Complementing Lexical Retrieval with Semantic Residual Embedding. arXiv:2004.13969 [cs.IR] https://arxiv.org/abs/2004.13969

[9] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling. arXiv:2104.06967 [cs.IR] https://arxiv.org/abs/2104.06967

[10] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised Dense Information Retrieval with Contrastive Learning. arXiv:2112.09118 [cs.IR] https://arxiv.org/abs/2112.09118

[11] Nasreen Jaleel, James Allan, W. Croft, Fernando Diaz, Leah Larkey, Xiaoyan Li, Mark Smucker, and Courtney Wade. 2004. UMass at TREC 2004: Novelty and hard.

[12] K Sparck Jones, Steve Walker, and Stephen E. Robertson. 2000. A probabilistic model of information retrieval: development and comparative experiments: Part 2. *Information processing & management* 36, 6 (2000), 809–840.

[13] Saar Kuzi, Mingyang Zhang, Cheng Li, Michael Bendersky, and Marc Najork. 2020. Leveraging Semantic and Lexical Matching to Improve the Recall of Document Retrieval Systems: A Hybrid Approach. arXiv:2010.01195 [cs.IR] https://arxiv.org/abs/2010.01195

[14] Joon Ho Lee. 1997. Analyses of multiple evidence combination. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Philadelphia, Pennsylvania, USA) *(SIGIR '97).* Association for Computing Machinery, New York, NY, USA, 267–276. doi:10.1145/258525.258587

[15] Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, Dense, and Attentional Representations for Text Retrieval. arXiv:2005.00181 [cs.CL] https://arxiv.org/abs/2005.00181

[16] Ji Ma, Ivan Korotkov, Keith B. Hall, and Ryan T. McDonald. 2020. Hybrid First-stage Retrieval Models for Biomedical Literature. In *Conference and Labs of the Evaluation Forum.* https://api.semanticscholar.org/CorpusID:221668044

[17] Mark Montague and Javed A. Aslam. 2001. Relevance score normalization for metasearch. In *Proceedings of the Tenth International Conference on Information and Knowledge Management* (Atlanta, Georgia, USA) *(CIKM '01).* Association for Computing Machinery, New York, NY, USA, 427–433. doi:10.1145/502585.502657

[18] Mark Montague and Javed A. Aslam. 2002. Condorcet fusion for improved retrieval. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management* (McLean, Virginia, USA) *(CIKM '02).* Association for Computing Machinery, New York, NY, USA, 538–548. doi:10.1145/584792.584881

[19] Giorgio Maria Di Nunzio, Stefano Marchesin 0001, and Federica Vezzani. 2020. A Study on Reciprocal Ranking Fusion in Consumer Health Search. IMS UniPD ad CLEF eHealth 2020 Task 2. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020 (CEUR Workshop Proceedings, Vol. 2696)*, Linda Cappellato, Carsten Eickhoff, Nicola Ferro 0001, and Aurélie Névéol (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-2696/paper_128.pdf

[20] Gerard Salton and Clement T Yu. 1973. On the construction of effective vocabularies for information retrieval. *Acm Sigplan Notices* 10, 1 (1973), 48–60.

[21] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and Permuted Pre-training for Language Understanding. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 16857–16867. https://proceedings.neurips.cc/paper_files/paper/2020/file/c3a690be93aa602ee2dc0ccab5b7b67e-Paper.pdf

[22] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).* https://openreview.net/forum?id=wCu6T5xFjeJ

[23] Shuai Wang, Shengyao Zhuang, and Guido Zuccon. 2021. BERT-based Dense Retrievers Require Interpolation with BM25 for Effective Passage Retrieval. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval* (Virtual Event, Canada) *(ICTIR '21).* Association for Computing Machinery, New York, NY, USA, 317–324. doi:10.1145/3471158.3472233

[24] Boya Zhang, Nona Naderi, Fernando Jaume-Santero, and Douglas Teodoro. 2022. DS4DH at TREC Health Misinformation 2021: Multi-Dimensional Ranking Models with Transfer Learning and Rank Fusion. arXiv:2202.06771 [cs.IR] https://arxiv.org/abs/2202.06771