Axiomatic Re-Ranking for Argument Retrieval

Maximilian Heinrich
maximilian.heinrich@uni-weimar.de
Bauhaus-Universität Weimar
Weimar, Germany

Marvin Vogel
mv15lyna@studserv.uni-leipzig.de
Leipzig University
Leipzig, Germany

Alexander Bondarenko
alexander.bondarenko@uni-jena.de
Friedrich-Schiller-Universität Jena
Jena, Germany

Matthias Hagen
matthias.hagen@uni-jena.de
Friedrich-Schiller-Universität Jena
Jena, Germany

Benno Stein
benno.stein@uni-weimar.de
Bauhaus-Universität Weimar
Weimar, Germany

Abstract

Information retrieval axioms are formalized constraints that retrieval systems should ideally satisfy (e.g., to rank documents higher that contain the query terms more often). In this paper, we propose new axioms that focus on the scenario of argument retrieval: retrieval for queries that need arguments in the results. Our underlying axiomatic idea is that in such scenarios, documents should be prioritized with argumentative units that are similar to the query. We test our new axioms in re-ranking experiments on the data of the Touché 2020 and 2021 shared task on argument retrieval for controversial questions, and show that the new axioms can improve the effectiveness of Touché's strong DirichletLM baseline model and even of the top-performing system from Touché 2021, a system already specifically optimized for argument retrieval. Finally, we also propose a new method for visualizing the relationships between axioms based on their effects in re-ranking settings.

CCS Concepts

• **Information systems** → **Retrieval models and ranking**; **Retrieval effectiveness**.

Keywords

Argument retrieval, Argument search, Axiomatic thinking for information retrieval, Axiomatic re-ranking

ACM Reference Format:
Maximilian Heinrich, Marvin Vogel, Alexander Bondarenko, Matthias Hagen, and Benno Stein. 2025. Axiomatic Re-Ranking for Argument Retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3726302.3730169

1 Introduction

The task of argument retrieval is to identify good arguments on a given topic. One use case is to support people who want to form an opinion on some controversial topic by looking for arguments that justify or refute some standpoint [27]. Interestingly, modern neural retrieval models seem to be inadequate for argument retrieval [26]—they are often much less effective than BM25. One reason could be that most retrieval models are argument-"agnostic" in the sense that they do not treat argumentative queries in a special way. In this paper, we thus propose to incorporate argumentation awareness into any retrieval model.

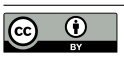
In particular, we follow an earlier idea of re-ranking some baseline model's retrieval results by using combinations of retrieval axioms [16]. Axiomatic thinking in information retrieval (IR) tries to capture basic important properties of good retrieval results in so-called "axioms", often formalized as constraints that induce document preferences (e.g., to prefer documents with more query term occurrences). For the scenario of argument retrieval, we propose new axioms that basically state to favor documents whose argumentative units are more similar to the query.

In an empirical evaluation on the datasets of the Touché 2020 and 2021 shared task on argument retrieval for controversial questions [4, 6], we show that re-ranking using the new argument-oriented axioms can improve the retrieval effectiveness of the shared task's strong DirichletLM baseline but also even of the best system submitted to the 2021 edition of the task. For the analysis, we also propose a novel approach to visualize axiom similarity via the axioms' impact on the re-ranking.[1]

2 Related Work

In the early days of axiomatic thinking in IR, such axioms were mainly meant to compare and improve retrieval models [8]. For instance, Fang et al. [11] presented a set of six retrieval "axioms" that relate to constraints on term frequency and document length that good retrieval results should fulfill (e.g., to favor documents with more query term occurrences). Later, also axioms focusing on, for instance, semantic similarity [13], proximity [25], or frequency normalization [18] were formalized and used to analyze retrieval models. Later studies then also used axioms to diversify retrieval results [15], to diagnose retrieval functions [12], to combine the results of multiple retrieval models [3], or to explain or improve neural rankings [9, 14, 19, 24, 28]. A use case that inspired us most are the studies that applied axioms in the re-ranking of retrieval results [7, 16] and that demonstrated that re-ranking some baseline models' results (e.g., BM25 [22, 23] or DirichletLM [30]) to better follow axiomatic preferences yields better retrieval effectiveness.

[1] All our code is available at: https://github.com/webis-de/SIGIR-25



SIGIR '25, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730169

3 Argumentation-Oriented Retrieval Axioms

Most retrieval axioms are formalized in a way that induces a ranking preference for pairs of documents. The formalization often consists of a precondition (e.g., the axiom is only applicable to documents of similar length), a filter condition (e.g., query term frequency), and a conclusion often formulated as an induced preference (e.g., from two same-length documents, the one with higher query term frequency should be ranked better). Targeting argument retrieval, we refine the STMC1 axiom [13]. In its original form, STMC1 favors documents whose individual terms are more similar to the query terms. But as arguments in documents do not occur on the level of individual terms but longer document units, we focus on the argumentative units of a document. In the two new QArgSim axioms, we formalize to prefer the document from a pair whose argumentative units on average are more similar to the query (the QArgSim$_{avg}$ axiom) or the document whose most similar argumentative unit is more similar to the query (the QArgSim$_{max}$). As argumentative units might not be that easy to detect but often span complete sentences, we also formalize sentence-level axioms analogous to the QArgSim axioms: QSenSim$_{avg}$ prefers the document whose sentences on average are more similar to the query, QSenSim$_{max}$ prefers the document whose most similar sentence is more similar to the query. Table 1 places the axiom ideas side by side.

None of our new axioms has preconditions but for each group, we define two variants: an exact variant (indicated by a superscript e) that expresses a preference for any similarity difference and a "relaxed" variant (no superscript) that outputs no preference if the documents' similarity scores are not substantially different (within a range of 10%). Thus, for instance, QArgSim$^{e}_{max}$ always expresses a preference, regardless of how small the actual difference between the similarity scores might be.

4 Experimental Evaluation

We assess the effectiveness of the new axioms within an axiomatic re-ranking scenario of the top-10 results of some basis retrieval model following the setup of Hagen et al. [16]: for instance, possible aggregation conflicts in the axiomatic pairwise preferences are handled by KwikSort [1]. Our experiments are conducted on the Touché 2020 and 2021 subtask of 'Argument Retrieval for Controversial Questions' [4, 6], that are based on the args.me corpus [2]. This corpus consists of about 390,000 argumentative documents from various debate portals and websites. For our experiments, we performed data cleaning and removed documents shorter than 15 characters and documents that only contain special characters—about 360,000 documents remained. The topics/queries in the shared task are controversial questions like 'Should plastic bottles be banned?' and relevance judgments were collected for the top-5 result pools of the submitted runs and a DirichletLM baseline. Thus, in our top-10 re-ranking setup we might encounter unjudged documents so that we decided to exclude unjudged documents from the retrieval pipeline while preserving the original ranking order of the retrieval models. Additionally, we exclude queries for which a given basis retriever returns fewer than 10 judged documents. For Touché 2020, we thus use a set of 44 queries with 2,002 judgments, while for Touché 2021 we can use all 50 queries with 3,654 judgments. To assess differences between an original and a re-ranked result set, we perform a paired t-test with Bonferroni correction ($\alpha = 0.05$). All experiments are implemented using the PyTerrier [20] and `ir_axioms` [5] toolkits.

4.1 Re-Ranking the DirichletLM Baseline

In a first experiment, we demonstrate the effectiveness of the argumentative axioms by re-ranking the top-10 results retrieved by the Touché tasks' baseline DirichletLM model. To identify argumentative units for the QArgSim axioms, we use the TARGER argument tagger [10] and view document segments as argumentative units that TARGER labels as either premise or claim. To assess the similarity between queries and argumentative units or sentences, we use Sentence-BERT (SBERT) embeddings [21] and compute cosine similarity [17]. As a baseline for comparison, we use the 38 traditional retrieval axioms implemented in the `ir_axioms` framework but reimplement the traditional STMC1 axiom using SBERT embeddings for a fair comparison. The nDCG@5 and nDCG@10 effectiveness of the re-rankings are shown in Table 2. For space reasons, only the three traditional axioms with the highest scores for each nDCG metric and each dataset are shown, resulting in a total of seven traditional axioms: RS-{retrieval_model} and M_AND [5], AND [29], PROX3 [16], and QTPArg [7].

The new argumentation-oriented axioms consistently achieve the highest scores among all tested axioms. This trend holds for both nDCG@5 and nDCG@10 across both datasets. However, none of the axioms achieves statistically significant improvements, likely due to the conservative nature of the Bonferroni correction. Still, our findings indicate that the new axioms are effective for argument retrieval. However, as the sentence-oriented axioms yield more effective retrieval results, the argumentative unit detection based on TARGER might have to be improved by using better detection models. Interestingly, the STMC1 axiom, which measures similarity at the term level and formed the basis for our new axioms, is not among the best-performing traditional axioms. Thus, our focus on longer document units seems to actually better capture the nature of argument retrieval. Moreover, the max-versions of the argumentative axioms tend to outperform the avg-versions. This may be attributed to the structure of the datasets, which contain texts from online debates, such as direct responses to opponents. By choosing the most similar sentence, the max-version can filter out irrelevant content like thank-you notes addressing the opponent, etc.

4.2 Re-ranking Touché 2021 Participants

In a second experiment, we demonstrate that the axioms can even improve the effectiveness of existing dedicated argument retrieval systems. To this end, we re-rank the best performing runs of the Touché 2021 participants who managed to beat the task's DirichletLM baseline. As in our first experiment, we re-rank the top-10 results. Table 3 shows the nDCG@5 and @10 scores as well as information on the overall ranking. Note that due to our slight cleaning modifications to the dataset, the participants' results may differ marginally from those reported in the original shared task.

Our results demonstrate that the participants would have benefited from applying the argumentative axioms on top of their retrieval systems. The improvement for several systems is statistically

Table 1: Overview of the four new axioms. The QArgSim axioms focus on the similarity of query and argumentative units of a document, while the QSenSim axioms focus on the similarity of query and sentences in a document.

QArgSim$_{avg}$	Prefer the document whose argument units on average are more similar to the query.
QArgSim$_{max}$	Prefer the document that contains the argument unit that is the most similar to the query.
QSenSim$_{avg}$	Prefer the document whose sentences on average are more similar to the query.
QSenSim$_{max}$	Prefer the document that contains the sentence that is the most similar to the query.

Table 2: Results for nDCG@5 and @10 scores after re-ranking the top 10 results for Touché 2020 (left) and Touché 2021 (right). The axioms are sorted according to the nDCG@5 score. Details about the axioms can be found in Section 4.1.

Touché 2020

	nDCG		Rank	
Axiom	**@5**	**@10**	**@5**	**@10**
QSenSim$^{e}_{max}$	0.813	0.775	1	1
QSenSim$_{max}$	0.795	0.769	2	2
QSenSim$_{avg}$	0.781	0.76	3	4
QArgSim$^{e}_{max}$	0.775	0.76	4	3
QSenSim$^{e}_{avg}$	0.772	0.76	5	5
QArgSim$_{avg}$	0.765	0.757	6	6
QArgSim$_{max}$	0.763	0.756	7	7
QArgSim$^{e}_{avg}$	0.757	0.754	8	8
AND	0.755	0.748	9	10
M_AND	0.753	0.749	10	9
RS_TF_IDF	0.746	0.748	11	11
DirichletLM	0.74	0.742	21	27
QTPArg	0.738	0.741	32	29
PROX3	0.734	0.737	36	41
RS_PL2	0.72	0.736	42	42
RS_TF	0.689	0.718	46	46

Touché 2021

	nDCG		Rank	
Axiom	**@5**	**@10**	**@5**	**@10**
QSenSim$^{e}_{avg}$	0.725	0.683	1	2
QSenSim$_{max}$	0.724	0.68	2	4
QSenSim$^{e}_{max}$	0.724	0.685	3	1
QSenSim$_{avg}$	0.717	0.682	4	3
QArgSim$^{e}_{max}$	0.694	0.671	5	5
QArgSim$_{max}$	0.687	0.67	6	6
QArgSim$_{avg}$	0.682	0.663	7	7
QArgSim$^{e}_{avg}$	0.677	0.659	8	8
RS_TF	0.661	0.658	9	9
QTPArg	0.66	0.652	10	12
PROX3	0.654	0.652	11	11
RS_PL2	0.653	0.654	18	10
DirichletLM	0.652	0.649	25	24
RS_TF_IDF	0.642	0.648	35	34
M_AND	0.63	0.645	44	41
AND	0.63	0.643	45	44

significant and re-ranked approaches from the middle range would have ranked among the top systems in the original shared task, while the already strong retrieval systems would have achieved even better results—except for the Asterix system.

Again, we observe that the max-version of QSenSim axioms yields the best results which shows that even a single axiom can significantly improve even argumentation-focused retrieval models.

4.3 Axiom Similarity

To further gain deeper insights into axiom similarity by comparisons of how they affect the re-ranking, we introduce a novel visualization method. The application of an axiom in re-ranking the results of a basis retrieval system leads to a change in the corresponding nDCG score. By computing this difference for each of k retrieval models (e.g., the ones from our experiment in Section 4.2), we derive a k-dimensional vector of axiom-specific difference scores. We take the cosine similarity between any two axioms' vectors to compute their distance in the k-dimensional space. The resulting distance matrix describes how close the effects of the axioms are for different retrieval systems. This proximity can be visualized by applying multidimensional scaling to generate a 2D plot (see Table 3, right). For better interpretation, we also include the effect size of the axioms as follows: summing the difference score vectors for each axiom and setting negative values to zero. In this way, the effect size of an axiom represents the total positive change in scores across all retrieval models.

The visualization reveals that the exact and non-exact versions of the arguments are closely related with only minimal differences. Additionally, distinct clusters are distinguishable for the sentence-based and for the argumentative unit-based axioms, as well as for their mean and max variants, indicating the respective axioms' comparable influence on re-ranking effectiveness.

5 Conclusion

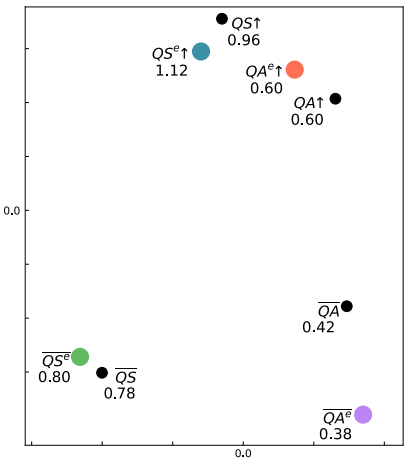
In this paper, we introduced four axioms for argument retrieval. The QArgSim axioms state to prefer documents where the argumentative units are more similar to the query, while the QSenSim axioms emphasize the similarity between the query and individual sentences. Our results of re-ranking experiments on the Touché 2020 and 2021 shared task of 'Argument Retrieval for Controversial

Table 3:
(Left) Results of the top-10 re-ranking of best retrieval systems submitted to Touché 2021. The *Orig* column shows the original score of the system, without axioms. The columns *QS* and *QA* (short for QSenSim and QArgSim) show the score if the participant's retrieval system was re-ranked with the corresponding axiom (overline: avg-version, upward arrow (↑): max-version, superscript ^{e}: exact version). (†) indicates a significant improvement compared to *Orig*. The 'Rank' column next to an axiom shows the re-ranked system's overall rank in the shared task, and in parentheses the rank gain. Results for nDCG@5 and nDCG@10 are shown in the upper and lower table respectively.
(Right) The plots illustrate the axiom similarity wrt. their effect on the different retrieval systems (see Section 4.3); the number under the name of an axiom indicates the average size of the effect (i.e., the "axiom impact").

nDCG@5 - Touché 2021

Participant	*Orig*	$\overline{QS^e}\uparrow$	Rank	$\overline{QS^e}$	Rank	$\overline{QA^e}\uparrow$	Rank	$\overline{QA^e}$	Rank
Elrond	0.72	0.792	1 (0)	**0.764**	1 (0)	0.761	1 (0)	0.737	1 (0)
Took	0.705	**0.797**†	1 (1)	0.762	1 (1)	**0.767**	1 (1)	**0.747**	1 (1)
Asterix	0.681	0.714	2 (1)	0.666	6 (-3)	0.701	3 (0)	0.696	3 (0)
Roberts	0.681	0.76†	1 (3)	0.75	1 (3)	0.686	3 (1)	0.686	3 (1)
Hood	0.668	0.738	1 (4)	0.684	3 (2)	0.725	1 (4)	0.73	1 (4)
Skeletor	0.668	0.74†	1 (5)	0.727	1 (5)	0.722	1 (5)	0.682	3 (3)
Shanks	0.66	0.76†	1 (6)	0.685	3 (4)	0.704	3 (4)	0.683	3 (4)
Skywalker	0.657	0.75†	1 (7)	0.742†	1 (7)	0.711	2 (6)	0.682	3 (5)
Deadpool	0.647	0.702	3 (6)	0.692	3 (6)	0.683	3 (6)	0.674	5 (4)
Heimdall	0.646	0.762†	1 (9)	0.75†	1 (9)	0.726	1 (9)	0.704	3 (7)
Athos	0.636	0.722†	1 (10)	0.705†	2 (9)	0.673	5 (6)	0.654	9 (2)
Ishikawa	0.634	0.704†	3 (9)	0.706†	2 (10)	0.672	5 (7)	0.654	9 (3)
Polnareff	0.634	0.726†	1 (12)	0.704†	3 (10)	0.663	7 (6)	0.655	9 (4)
DirichletLM	0.626	0.718†	2 (12)	0.712†	2 (12)	0.665	7 (7)	0.659	8 (6)

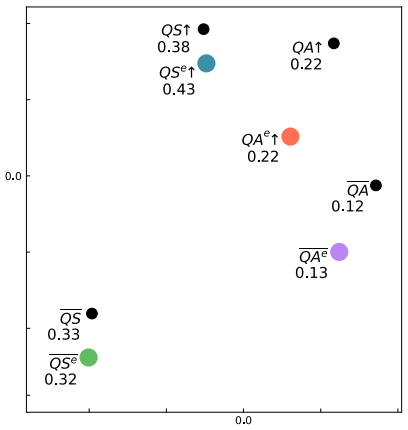
Axiom Similarity and Impact (nDCG@5)



nDCG@10 - Touché 2021

Participant	*Orig*	$\overline{QS^e}\uparrow$	Rank	$\overline{QS^e}$	Rank	$\overline{QA^e}\uparrow$	Rank	$\overline{QA^e}$	Rank
Elrond	0.718	0.747	1 (0)	0.733	1 (0)	0.733	1 (0)	0.724	1 (0)
Took	0.717	**0.756**†	1 (1)	**0.737**	1 (1)	**0.742**	1 (1)	**0.735**	1 (1)
Roberts	0.685	0.715	3 (0)	0.712	3 (0)	0.685	3 (0)	0.685	3 (0)
Heimdall	0.673	0.713†	3 (1)	0.712	3 (1)	0.697	3 (1)	0.689	3 (1)
Skeletor	0.672	0.708†	3 (2)	0.697	3 (2)	0.696	3 (2)	0.684	4 (1)
Hood	0.668	0.687	3 (3)	0.667	6 (0)	0.681	4 (2)	0.682	4 (2)
Skywalker	0.667	0.697	3 (4)	0.693	3 (4)	0.687	3 (4)	0.672	6 (1)
Asterix	0.667	0.674	4 (4)	0.656	9 (-1)	0.67	6 (2)	0.67	6 (2)
Shanks	0.662	0.699†	3 (6)	0.67	6 (3)	0.676	4 (5)	0.669	6 (3)
Deadpool	0.644	0.67	6 (4)	0.669	6 (4)	0.662	9 (1)	0.655	10 (0)
Athos	0.638	0.673†	4 (7)	0.674†	4 (7)	0.656	10 (1)	0.648	10 (1)
Polnareff	0.634	0.671†	6 (6)	0.665†	9 (3)	0.646	10 (2)	0.643	11 (1)
DirichletLM	0.633	0.669	6 (7)	0.67†	6 (7)	0.649	10 (3)	0.647	10 (3)
Ishikawa	0.63	0.66†	10 (4)	0.665†	9 (5)	0.648	10 (4)	0.641	11 (3)

Axiom Similarity and Impact (nDCG@10)



Questions' show that the QSenSim axioms consistently outperform other axiom groups, yielding significant improvements in nDCG@5 and @10 scores. In particular the 'max' variants typically perform well, probably due to their ability to focus only on the most relevant parts of a document. With new visualizations of axiom effects when re-ranking some existing retrieval models' results, we are further able to identify similar axioms. An interesting direction for future research is to apply the new "argumentative" axioms also in non-argumentative contexts.

Acknowledgments

This work was supported by the German Federal Ministry of Education and Research (BMBF) through the project "DIALOKIA: Überprüfung von LLM-generierter Argumentation mittels dialektischem Sprachmodell" (01IS24084A-B), and by the German Research Foundation (DFG) through the project "ACQuA 2.0: Answering Comparative Questions with Arguments" (376430233).

References

[1] Nir Ailon, Moses Charikar, and Alantha Newman. 2008. Aggregating Inconsistent Information: Ranking and Clustering. *J. ACM* 55, 5 (2008), 23:1–23:27. doi:10.1145/1411509.1411513

[2] Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Data Acquisition for Argument Search: The args.me Corpus. In *KI 2019: Advances in Artificial Intelligence - 42nd German Conference on AI, Kassel, Germany, September 23-26, 2019, Proceedings (Lecture Notes in Computer Science, Vol. 11793)*, Christoph Benzmüller and Heiner Stuckenschmidt (Eds.). Springer, 48–59. doi:10.1007/978-3-030-30179-8_4

[3] Siddhant Arora and Andrew Yates. 2019. Investigating Retrieval Method Selection with Axiomatic Features. In *Proceedings of the 1st Interdisciplinary Workshop on Algorithm Selection and Meta-Learning in Information Retrieval co-located with the 41st European Conference on Information Retrieval (ECIR 2019), Cologne, Germany, April 14, 2019 (CEUR Workshop Proceedings, Vol. 2360)*, Jöran Beel and Lars Kotthoff (Eds.). CEUR-WS.org, 18–31. https://ceur-ws.org/Vol-2360/paper4Axiomatic.pdf

[4] Alexander Bondarenko, Maik Fröbe, Meriem Beloucif, Lukas Gienapp, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2020. Overview of Touché 2020: Argument Retrieval – Extended Abstract. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22-25, 2020, Proceedings (Lecture Notes in Computer Science, Vol. 12260)*, Avi Arampatzis, Evangelos Kanoulas, Theodora Tsikrika, Stefanos Vrochidis, Hideo Joho, Christina Lioma, Carsten Eickhoff, Aurélie Névéol, Linda Cappellato, and Nicola Ferro (Eds.). Springer, 384–395. doi:10.1007/978-3-030-58219-7_26

[5] Alexander Bondarenko, Maik Fröbe, Jan Heinrich Reimer, Benno Stein, Michael Völske, and Matthias Hagen. 2022. Axiomatic Retrieval Experimentation with ir_axioms. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai (Eds.). ACM, 3131–3140. doi:10.1145/3477495.3531743

[6] Alexander Bondarenko, Lukas Gienapp, Maik Fröbe, Meriem Beloucif, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2021. Overview of Touché 2021: Argument Retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21-24, 2021, Proceedings (Lecture Notes in Computer Science, Vol. 12880)*, K. Selçuk Candan, Bogdan Ionescu, Lorraine Goeuriot, Birger Larsen, Henning Müller, Alexis Joly, Maria Maistro, Florina Piroi, Guglielmo Faggioli, and Nicola Ferro (Eds.). Springer, 450–467. doi:10.1007/978-3-030-85251-1_28

[7] Alexander Bondarenko, Matthias Hagen, Michael Völske, Benno Stein, Alexander Panchenko, and Chris Biemann. 2018. Webis at TREC 2018: Common Core Track. In *Proceedings of the Twenty-Seventh Text REtrieval Conference, TREC 2018, Gaithersburg, Maryland, USA, November 14-16, 2018 (NIST Special Publication, Vol. 500-331)*, Ellen M. Voorhees and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec27/papers/Webis-CC.pdf

[8] Peter Bruza and Theo W. C. Huibers. 1994. Investigating Aboutness Axioms using Information Fields. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland, 3-6 July 1994 (Special Issue of the SIGIR Forum)*, W. Bruce Croft and C. J. van Rijsbergen (Eds.). ACM/Springer, 112–121. doi:10.1007/978-1-4471-2099-5_12

[9] Arthur Câmara and Claudia Hauff. 2020. Diagnosing BERT with Retrieval Heuristics. In *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 12035)*, Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins (Eds.). Springer, 605–618. doi:10.1007/978-3-030-45439-5_40

[10] Artem N. Chernodub, Oleksiy Oliynyk, Philipp Heidenreich, Alexander Bondarenko, Matthias Hagen, Chris Biemann, and Alexander Panchenko. 2019. TARGER: Neural Argument Mining at Your Fingertips. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 3: System Demonstrations*, Marta R. Costa-jussà and Enrique Alfonseca (Eds.). Association for Computational Linguistics, 195–200. doi:10.18653/V1/P19-3031

[11] Hui Fang, Tao Tao, and ChengXiang Zhai. 2004. A Formal Study of Information Retrieval Heuristics. In *SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, July 25-29, 2004*, Mark Sanderson, Kalervo Järvelin, James Allan, and Peter Bruza (Eds.). ACM, 49–56. doi:10.1145/1008992.1009004

[12] Hui Fang, Tao Tao, and ChengXiang Zhai. 2011. Diagnostic Evaluation of Information Retrieval Models. *ACM Trans. Inf. Syst.* 29, 2 (2011), 7:1–7:42. doi:10.1145/1961209.1961210

[13] Hui Fang and ChengXiang Zhai. 2006. Semantic Term Matching in Axiomatic Approaches to Information Retrieval. In *SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, August 6-11, 2006*, Efthimis N. Efthimiadis, Susan T. Dumais, David Hawking, and Kalervo Järvelin (Eds.). ACM, 115–122. doi:10.1145/1148170.1148193

[14] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. A White Box Analysis of ColBERT. In *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 12657)*, Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani (Eds.). Springer, 257–263. doi:10.1007/978-3-030-72240-1_23

[15] Sreenivas Gollapudi and Aneesh Sharma. 2009. An Axiomatic Approach for Result Diversification. In *Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009*, Juan Quemada, Gonzalo León, Yoëlle S. Maarek, and Wolfgang Nejdl (Eds.). ACM, 381–390. doi:10.1145/1526709.1526761

[16] Matthias Hagen, Michael Völske, Steve Göring, and Benno Stein. 2016. Axiomatic Result Re-Ranking. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, Snehasis Mukhopadhyay, ChengXiang Zhai, Elisa Bertino, Fabio Crestani, Javed Mostafa, Jie Tang, Luo Si, Xiaofang Zhou, Yi Chang, Yunyao Li, and Parikshit Sondhi (Eds.). ACM, 721–730. doi:10.1145/2983323.2983704

[17] Baoli Li and Liping Han. 2013. Distance Weighted Cosine Similarity Measure for Text Classification. In *Intelligent Data Engineering and Automated Learning - IDEAL 2013 - 14th International Conference, IDEAL 2013, Hefei, China, October 20-23, 2013. Proceedings (Lecture Notes in Computer Science, Vol. 8206)*, Hujun Yin, Ke Tang, Yang Gao, Frank Klawonn, Minho Lee, Thomas Weise, Bin Li, and Xin Yao (Eds.). Springer, 611–618. doi:10.1007/978-3-642-41278-3_74

[18] Yuanhua Lv and ChengXiang Zhai. 2011. Lower-Bounding Term Frequency Normalization. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*, Craig Macdonald, Iadh Ounis, and Ian Ruthven (Eds.). ACM, 7–16. doi:10.1145/2063576.2063584

[19] Sean MacAvaney, Sergey Feldman, Nazli Goharian, Doug Downey, and Arman Cohan. 2022. ABNIRML: Analyzing the Behavior of Neural IR Models. *Trans. Assoc. Comput. Linguistics* 10 (2022), 224–239. doi:10.1162/TACL_A_00457

[20] Craig Macdonald and Nicola Tonellotto. 2020. Declarative Experimentation in Information Retrieval using PyTerrier. In *ICTIR '20: The 2020 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Norway, September 14-17, 2020*, Krisztian Balog, Vinay Setty, Christina Lioma, Yiqun Liu, Min Zhang, and Klaus Berberich (Eds.). ACM, 161–168. doi:10.1145/3409256.3409829

[21] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 3980–3990. doi:10.18653/V1/D19-1410

[22] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994 (NIST Special Publication, Vol. 500-225)*, Donna K. Harman (Ed.). National Institute of Standards and Technology (NIST), 109–126. http://trec.nist.gov/pubs/trec3/papers/city.ps.gz

[23] Stephen E. Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* 3, 4 (2009), 333–389. doi:10.1561/1500000019

[24] Corby Rosset, Bhaskar Mitra, Chenyan Xiong, Nick Craswell, Xia Song, and Saurabh Tiwary. 2019. An Axiomatic Approach to Regularizing Neural Ranking Models. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer (Eds.). ACM, 981–984. doi:10.1145/3331184.3331296

[25] Tao Tao and ChengXiang Zhai. 2007. An Exploration of Proximity Measures in Information Retrieval. In *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*, Wessel Kraaij, Arjen P. de Vries, Charles L. A. Clarke, Norbert Fuhr, and Noriko Kando (Eds.). ACM, 295–302. doi:10.1145/1277741.1277794

[26] Nandan Thakur, Luiz Bonifacio, Maik Fröbe, Alexander Bondarenko, Ehsan Kamalloo, Martin Potthast, Matthias Hagen, and Jimmy Lin. 2024. Systematic Evaluation of Neural Retrieval Models on the Touché 2020 Argument Retrieval Subset of BEIR. 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*, Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang (Eds.). ACM, 1420–1430. doi:10.1145/3626772.3657861

[27] Frans H. van Eemeren and Rob Grootendorst. 2003. *A Systematic Theory of Argumentation: The Pragma-Dialectical Approach.* Cambridge University Press, Cambridge. doi:10.1017/CBO9780511616389

[28] Michael Völske, Alexander Bondarenko, Maik Fröbe, Benno Stein, Jaspreet Singh, Matthias Hagen, and Avishek Anand. 2021. Towards Axiomatic Explanations for Neural Ranking Models. In *ICTIR '21: The 2021 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Canada, July 11, 2021*, Faegheh Hasibi, Yi Fang, and Akiko Aizawa (Eds.). ACM, 13–22. doi:10.1145/3471158.3472256

[29] Hao Wu and Hui Fang. 2012. Relation Based Term Weighting Regularization. In *Advances in Information Retrieval - 34th European Conference on IR Research, ECIR 2012, Barcelona, Spain, April 1-5, 2012. Proceedings (Lecture Notes in Computer Science, Vol. 7224)*, Ricardo Baeza-Yates, Arjen P. de Vries, Hugo Zaragoza, Berkant Barla Cambazoglu, Vanessa Murdock, Ronny Lempel, and Fabrizio Silvestri (Eds.). Springer, 109–120. doi:10.1007/978-3-642-28997-2_10

[30] ChengXiang Zhai and John D. Lafferty. 2001. A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. In *SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, September 9-13, 2001, New Orleans, Louisiana, USA*, W. Bruce Croft, David J. Harper, Donald H. Kraft, and Justin Zobel (Eds.). ACM, 334–342. doi:10.1145/383952.384019