The 34th Canadian Conference on Artificial Intelligence
DOI: 0

Canadian AI 2021
(Vancouver, May 25-28, 2021)

Document Specificity Measures for Ad-hoc Retrieval

Abstract

When searching, users are interested in accessing the most *relevant* and *specific* content related to their information need. Earlier research has shown that it is much easier to retrieve appropriate content for specific queries compared to generic ones as it is possible to discriminatively distinguish the content related to specific queries. The work in this paper builds on earlier findings on query and document specificity and provides a systematic account of ways through which document specificity can be measured. We present a comprehensive view of how various measures of document specificity can be defined and comparatively analyze the utility of various document specificity measures within the context of ad hoc retrieval based on three well-known TREC corpora, namely Robust04, ClueWeb09B, ClueWeb12B and their associated TREC topics. We report on our findings on the effectiveness of each type of document specificity measure.

Keywords: Document specificity, ad hoc retrieval, document reranking

1. Introduction

Ad hoc retrieval is concerned with the effective ranking of documents given a query. Existing works have shown that the effectiveness of retrieval models varies across different query types leading to good performance on some queries (soft queries) and poor performance on others (hard queries) [1]. When exploring the characteristics of queries, it is possible to see that the performance of queries is *not* necessarily only dependent on the number of relevant documents retrieved for each query. In other words, harder queries are not always those queries that the retrieval models have a hard time finding relevant documents for, but can also be queries for which the retrieval models fail to effectively rank the relevant documents.

Earlier research on query performance shows that performance is correlated with query *specificity* where generic queries are harder and specific queries are softer [2]. The main reason for this is that the more specific a query is, the less likely it would be for it to share similarity with irrelevant documents and as such better results are obtained for such a query. The above two observations, i.e., (i) the lack of exclusive correlation between the number of relevant retrieved documents and query performance; and (ii) the relationship between query performance and *specificity*, is aligned with existing work in the literature, which have shown that the consideration of *document specificity* can lead to improved retrieval performance [3–6]. However, these works only consider a limited type of document specificity focused on inter-document associations captured through language models. There are, to the best of our knowledge, little, if any work, that considers (a) *document content*, and (b) *neural embeddings* when measuring specificity. As such, the **objective of our work** in this paper is (1) to systematically collect, introduce and classify document specificity metrics from both perspectives of *structure-based specificity* and *content-based specificity*, (2) to understand the impact of each of these types of specificity on improving the performance of document retrieval, and (3) study the possible synergistic impact of these specificity types on each other. Therefore, our work is the first to provide a holistic view of various types of document specificity metrics in the context of document retrieval, which systematically evaluates their performance on different standard TREC corpora. Based on the proposed classification of document specificity metrics, we answer three main Research Questions (RQs): (**RQ1**) Whether the consideration of structure-based specificity metrics have any impact on document retrieval? (**RQ2**) Would content-based document specificity metrics have noticeable impact on document retrieval? and (**RQ3**) Do structure-based document

This article is © 2024 by author(s) as listed above. The article is licensed under a Creative Commons Attribution (CC BY 4.0) International license (https://creativecommons.org/licenses/by/4.0/legalcode), except where otherwise indicated with respect to particular material included in the article. The article should be attributed to the author(s) identified above.

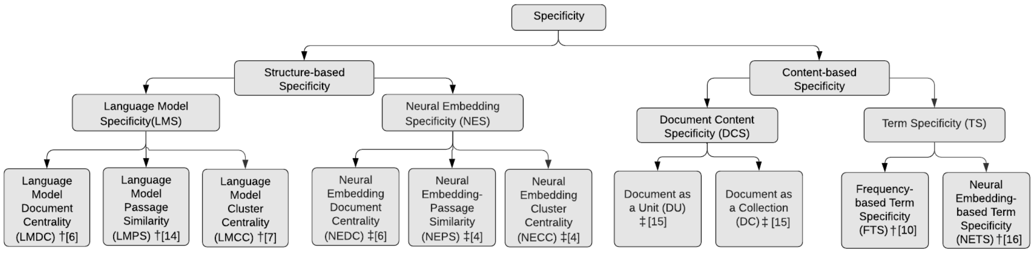
Specificity
Structure-based Specificity
Content-based Specificity
Language Model Specificity(LMS)
Neural Embedding Specificity (NES)
Document Content Specificity (DCS)
Term Specificity (TS)
Language Model Document Centrality (LMDC) †[6]
Language Model Passage Similarity (LMPS) †[14]
Language Model Cluster Centrality (LMCC) †[7]
Neural Embedding Document Centrality (NEDC) ‡[6]
Neural Embedding-Passage Similarity (NEPS) ‡[4]
Neural Embedding Cluster Centrality (NECC) ‡[4]
Document as a Unit (DU) ‡[15]
Document as a Collection (DC) ‡ [15]
Frequency-based Term Specificity (FTS) † [10]
Neural Embedding-based Term Specificity (NETS) † [16]

Figure 1. Classification of specificity measures. †shows method is introduced in reference, ‡shows method is *inspired by* reference yet only introduced in this paper.

specificity metrics and content-based document specificity metrics have synergistic impact on each other for document retrieval?

We note that the objective of this paper is not to show that document specificity is able to show SoTA performance on document retrieval, but rather it intends to (a) provide a comprehensive view of document specificity by collecting and introducing both structure-based and content-based measures, and (b) comparatively analyze the performance of these measures to understand their utility in practice. We empirically report our findings on large-scale document collections, namely, ClueWeb09, ClueWeb12, and Robust04 based on NDCG, MAP, and Precision at ranks 5 and 10.

2. Measures of Specificity

We propose that *document specificity* can be viewed from two perspectives: *structure-based* and *content-based* as shown in Figure 1. The figure indicates which measures are directly adopted from the literature, and which are inspired by work in the literature yet only introduced in this paper.

2.1. Structure-based Specificity

Structure-based measures consider the association between the documents to determine the degree of specificity of each document. Specificity of a document can be defined depending on how it is related to other documents [6]. Such measures of specificity depend on inter-document associations that are often computed through Language Models [3–6]. Based on such associations, a network of documents are formed where each document acts as a node and the inter-document associations are the edges. Structure-based specificity for each document would then be computed based on its position in the network.

Language Model (LM) Structure-based Specificity: LM structure-based specificity measures can be categorized into three classes, namely Document Centrality, Passage Similarity, and Cluster Centrality.

Language Model Document Centrality (LMDC): One of the foundations of structural specificity measures is to compute the centrality of each document within a collection. Kurland et al. [3] have proposed that utilizing PageRank would be a suitable metric for measuring document centrality. Based on [3], it is possible to compute the PageRank structure-based specificity metric, denoted by $Score_{PR}(d)$, as follows: Given a set of documents, a weighted directed graph G can be formed where the vertices are the set of documents and $E_{\mathbb{G}} = \{(e_i, e_j, w) | e_i \in \text{Nbhd } (\text{e}_j, \delta)\}$ where $Nbhd(e_i, \delta)$ is the neighborhood of top-δ documents with the highest probability of generating e_j, i.e., $P_{e_i}(e_j)$ and w is defined as $P_{e_i}(e_j)$. Simply put, in this weighted graph, an edge exists from document e_i to document e_j if document e_i is among the top-δ documents with the highest probability to generate document e_j. On this basis, $Score_{PR}(d)$ is defined as the normalized PageRank centrality of d in $\mathbb{G}$. Here $P_{e_i}(e_j)$ is defined as the Kullback-Leiber Divergence between LMs of documents e_i and e_j.

Language Model Passage Similarity (LMPS): The literature suggests that information induced from passages can be strong sources for measuring specificity [7]. A passage is a short query-specific representation of a document that serves as a proxy for that document when retrieving documents for the query. Therefore, inspired by Krikon et al., we define $Score_{psg}(d, q)$ for document d, which constitutes half-overlapping fixed window passages g_i $\in$d$, as follows :$

$$\text{Score}_{psg}(d,q) = \lambda \frac{Score_{PR}(d)p_d(q)}{\sum_{d' \in D_{init}} Score_{PR}(d')p_{d'}(q)} + (1-\lambda)\frac{\max_{g_i \in d} p_{g_i}(q)Score_{PR}(g_i)}{\sum_{d' \in D_{init}} \max_{g' \in d'} p_{g'}(q)Score_{PR}(g')}$$

where λ is a free parameter, $\mathbb{D}_{init}$ is the retrieved list of documents, and $P_g(q)$ and $P_d(q)$ are defined based on the same method as LMDC. Unlike LMDC, given the fact that the extraction of a passage from the input text is dependent on a query, LMPS is also directly dependent on the input query q.

Language Model Cluster Centrality (LMCC): It has been widely argued that ambiguous queries can be interpreted in different ways [4]; as such, documents retrieved for such queries can be the reflection of the multiplicity of interpretations. Kurland et al have shown that Clustering methods can identify different senses of ambiguous queries where the cluster with the highest percentage of relevant documents can be defined as an *optimal-cluster*. Given the clusters, one can define centrality measures based on the similarity between document-cluster, document-query, and cluster-query associations. To this end, LMCC incorporates centrality among the clusters and documents. Given query q, for each document d, we consider cluster c of documents that includes d. $Score_{cluster}(c, q)$ is defined as follows, where $P_c(q)$, $P_d(q)$, and $P_{d_i}(c)$ are defined same as LMDC method:

$$\text{Score}_{cluster}(c,q) = \lambda Score_{PR}(c)p_c(q) + (1-\lambda)\sum_{d_i \in c} p_{d_i}(q)p_{d_i}(c)Score_{PR}(d_i)$$

Neural Embedding Structure-based Specificity: While structure-based specificity metrics are often defined based on an LM, we further propose to use neural embeddings to compute inter-document associations. We note that the definition of the three metrics in this class are similar to the three language model structure based specificity metrics, except for the fact that $P_{e_i}(ej)$ is defined as the Word Mover's Distance (WMD) [8] between e_i and e_j. Given the fact that neural representations are used to compute and define inter-document associations instead of using LMs, we refer to the counterparts of the language model based metrics, namely LMDC, LMPS, and LMCC as Neural Embedding Document Centrality **(NEDC)**, Neural Embedding Passage Similarity **(NEPS)**, and Neural Embedding Cluster Centrality **(NECC)**, respectively.

2.2. Content-based Specificity

This class of specificity measures focuses on the content value of each document rather than the inter-document associations, which can be computed at either the *document* (document content specificity) or the *term* (term specificity) levels.

Document Content Specificity: Document content specificity need to be measured based on the representation of the content of each document.It is possible to learn document representations either (1) by considering the representation of each individual term appearing in the document, i.e., *Document as a Collection* (DC), using the average of the representations of the terms in that document or (2) by learning a unique representation for each document, i.e., *Document as a Unit* (DU), using methods such as doc2vec [9] .To measure content specificity, we first learn representations for each document either as DC or DU and then measure specificity of the document. Among specificity metrics [10], we adopt Edge Weight Sum (EWS), as it is computationally inexpensive and has shown good performance in various IR tasks [11].

Term Specificity An alternative approach is to calculate the specificity of the document's constituting terms. One can define term metrics that operate based on either term frequency statistics in the document collection or the geometric properties of term representations within an embedding space.

Frequency-based Term Specificity metrics (FTS): For the purpose of computing term-based specificity according to term frequency, we adopt the well-known frequency-based specificity metric, called Inverse Document Frequency (IDF). The IDF metric has shown promising performance in different IR tasks, such as query performance prediction [12]. We adopt IDF as a frequency-based document content specificity metric that relies on the frequency of each individual term of the document across the corpus.

Neural Embedding-based Term Specificity metrics (NETS): It is also possible to employ the neural embedding-based specificity metrics for each term that appears in the document. Similar to document specificity metrics, we adopt EWS [10] in order to calculate the average specificity of all terms in a document.

3. Experiments

Experimental Setup: We used ClueWeb09B, which consists of the first 50 million English Web pages of ClueWeb09; ClueWeb12B, which is a subset of over 50 million documents from ClueWeb12; and Robust04 consisting of 528,155 documents. We used TREC topics related to each corpus. For ClueWeb09B, topics 1-200, for ClueWeb12B, topics 201-250, and for Robust04, topics 301-450 and 601-700 were used. For the ranking model, we used the widely adopted work by Metzler and Croft [13]. We used the runs publicly shared in [**eqfe**]. As suggested by [6], we set the initial list of retrieved documents to the top-50 documents retrieved by the ranking model. As embeddings, we used the pre-trained model on Google News [14]. It is important to note that Krikon et al have already shown that document retrieval based on structure-based specificity can improve results over PRF methods such as RM3; therefore, given space limitations, we do not report similar results for other re-rankers noting consistency of our results with [6]. Performance was evaluated with MAP, NDCG and Precision at 5 and 10. Statistical significance is based on paired t-test (95% confidence).
Retrieval Framework: To ranking documents, we jointly considered document relevance and specificity when ranking documents for a given query. Given a document D and a query Q, the score of the document for the query can be expressed as Eq.3, where λ is linear interpolation coefficients and $f_{ret}(D,Q)$ is the normalized value of the baseline ranking function that computes the relatedness of D for Q. λ is set by sweep as suggested in [3, 4, 7] from {0, 0.1,..,1}.

$$Score(D,Q) = \lambda \underbrace{f_{ret}(D,Q)}_{\text{Retrieval Model}} + (1-\lambda) \underbrace{f_{spec}(D,Q)}_{\text{Specificity Model}}$$

Results: We report the results of the retrieval process based on different document specificity metrics in Table 1 to answer our three research questions. **RQ1** investigates whether structure-based specificity metrics can improve the document retrieval task. Our experiments show two **noteworthy findings**: **(1)** those structure-based specificity measures that perform clustering, e.g., LMCC and NECC, are able to show a statistically significant improved performance on document retrieval. **(2)** structure-based neural embedding measures introduced in this paper are far more effective than their language model-based counterparts

Table 1. Retrieval performance on Clueweb09, ClueWeb12 and Robust04 . Statistical significance at 95% confidence interval is denoted by *.

	Category	Method	P@5		P@10		MAP@5		MAP@10		NDCG@5		NDCG@10	
			value	Δ%	value	Δ%	value	Δ%	value	Δ%	value	Δ%	value	Δ%
ClueWeb09 Collection	Language Model Structure-based	LMDC	0.393	0.51	0.371	0.54	0.022	15.79*	0.036	12.50*	0.061	8.93*	0.087	4.82*
		LMPS	0.398	1.79	0.372	0.81	0.022	15.79*	0.036	12.50*	0.059	5.36*	0.086	3.61
		LMCC	0.399	2.05	0.389	5.42*	0.022	15.79*	0.036	12.50*	0.061	8.93*	0.086	3.61
	Neural Embedding Structure-based	NEDC	0.419	7.16*	**0.390**	**5.69***	0.023	21.05*	**0.039**	**21.88***	**0.069**	**23.21***	**0.093**	**12.05***
		NEPS	0.398	1.80	0.372	0.81	0.022	15.79*	0.04	12.50*	0.590	5.35*	0.086	3.61
		NECC	0.411	5.12*	0.371	0.54	**0.025**	**31.58***	**0.039**	**21.88***	0.065	16.07*	0.092	10.84*
	Document Content Specificity	DU	0.399	2.05	0.371	0.54	0.020	5.26*	0.033	3.13	0.057	1.79	0.083	0.00
		DC	0.400	2.30	0.381	3.25*	0.020	4.21	0.036	12.50*	0.058	3.57*	0.086	3.61*
	Term Specificity	FTS	0.410	4.86*	0.387	4.88*	0.019	0.00	0.033	3.13	0.057	1.79	0.086	3.61
		NETS	**0.424**	**8.44***	**0.390**	**5.69***	0.021	10.53*	0.036	12.50*	0.060	7.14	0.087	4.82*
ClueWeb12 Collection	Language Model Structure-based	LMDC	0.393	0.51	0.371	0.54	0.022	15.79*	0.036	12.50*	0.061	8.93*	0.087	4.82*
		LMPS	0.398	1.79	0.372	0.81	0.022	15.79*	0.036	12.50*	0.059	5.36*	0.086	3.61
		LMCC	0.399	2.05	0.389	5.42*	0.022	15.79*	0.036	12.50*	0.061	8.93*	0.086	3.61
	Neural Embedding Structure-based	NEDC	0.419	7.16*	**0.390**	**5.69***	0.023	21.05*	**0.039**	**21.88***	**0.069**	**23.21***	**0.093**	**12.05***
		NEPS	0.398	1.80	0.372	0.81	0.022	15.79*	0.04	12.50*	0.590	5.35*	0.086	3.61
		NECC	0.411	5.12*	0.371	0.54	**0.025**	**31.58***	**0.039**	**21.88***	0.065	16.07*	0.092	10.84*
	Document Content Specificity	DU	0.399	2.05	0.371	0.54	0.020	5.26*	0.033	3.13	0.057	1.79	0.083	0.00
		DC	0.400	2.30	0.381	3.25*	0.020	4.21	0.036	12.50*	0.058	3.57*	0.086	3.61*
	Term Specificity	FTS	0.410	4.86*	0.387	4.88*	0.019	0.00	0.033	3.13	0.057	1.79	0.086	3.61
		NETS	**0.424**	**8.44***	**0.390**	**5.69***	0.021	10.53*	0.036	12.50*	0.060	7.14	0.087	4.82*
Robust04 Collection	Language Model Structure-based	LMDC	0.393	0.51	0.371	0.54	0.022	15.79*	0.036	12.50*	0.061	8.93*	0.087	4.82*
		LMPS	0.398	1.79	0.372	0.81	0.022	15.79*	0.036	12.50*	0.059	5.36*	0.086	3.61
		LMCC	0.399	2.05	0.389	5.42*	0.022	15.79*	0.036	12.50*	0.061	8.93*	0.086	3.61
	Neural Embedding Structure-based	NEDC	0.419	7.16*	**0.390**	**5.69***	0.023	21.05*	**0.039**	**21.88***	**0.069**	**23.21***	**0.093**	**12.05***
		NEPS	0.398	1.80	0.372	0.81	0.022	15.79*	0.04	12.50*	0.590	5.35*	0.086	3.61
		NECC	0.411	5.12*	0.371	0.54	**0.025**	**31.58***	**0.039**	**21.88***	0.065	16.07*	0.092	10.84*
	Document Content Specificity	DU	0.399	2.05	0.371	0.54	0.020	5.26*	0.033	3.13	0.057	1.79	0.083	0.00
		DC	0.400	2.30	0.381	3.25*	0.020	4.21	0.036	12.50*	0.058	3.57*	0.086	3.61*
	Term Specificity	FTS	0.410	4.86*	0.387	4.88*	0.019	0.00	0.033	3.13	0.057	1.79	0.086	3.61
		NETS	**0.424**	**8.44***	**0.390**	**5.69***	0.021	10.53*	0.036	12.50*	0.060	7.14	0.087	4.82*

Table 2. Percentage of queries improved by both content and structure measures.

Results on MAP@5		Content-based Specificity					
		CW09		CW12		Robust04	
		DU	NETS	DU	NETS	DU	NETS
Structure-based Specificity	NECC	19%	23%	21%	17%	29%	34%
	NEDC	68%	28%	23%	15%	56%	65%

for document retrieval. There are two reasons for this: (a) LMs have several free parameters that need to be optimized, which is not the case for embedding-based measures; and (b) embedding-based measures consider the semantic association between documents, which is not captured by LMs.

Now, in **RQ2**, we investigate whether content-based specificity metrics have any positive impact on document retrieval. The **important finding** of our experiments is that most content-based specificity metrics have limited impact on document retrieval. The performance of these measures are in contrast to structure-based specificity measures, which are quite strong for document retrieval. With this insight into the performance of structure-based and content-based methods, in **RQ3**, we are interested to see whether these two types of measures have any synergistic impact on each other or not. In order to investigate this synergy, we have selected two top performing metrics from each class, namely NECC and NEDC from the structure-based measures, and DU and NETS from the content-based measures. Table 2 compares two metrics in each class with each other and shows the percentage of shared queries that both improved. For instance, the table shows that the NEDC and DU measures only have a 19% overlap in terms of the queries that they had improved on ClueWeb09. In contrast, NECC and NETS have a high degree of overlap on the queries that they improved, i.e., 65%. The higher the degree of overlap between two measures, the more correlated and less synergistic they are. The **insightful finding** from Table 2 is that while content-based specificity metrics are not strong on their own for effective document retrieval, they show complementary behavior to structure-based measures. Hence, they have the potential to improve the overall performance of the document retrieval task if and when systematically interpolated with structure-based measures, especially NEDC. Therefore in response to RQ3, content-based measures are able to improve queries that could not be otherwise addressed by structure-based methods and so show synergistic impact.

4. Concluding Remarks

We have curated, introduced and classified document specificity measures in structure-based and content-based categories. From our experiments, we can draw **important and impactful conclusions**: **(1)** structure-based specificity metrics are successful in improving the retrieval process; **(2)** from structure-based measures, document associations computed based on neural embeddings are far more effective compared to those measured based on language models; **(3)** content-based specificity measures are not effective for document retrieval; yet **(4)** they exhibit synergistic behavior to structure-based measures; hence, have the potential to lead to stronger retrieval based on the interpolation of content-based and structure-based specificity measures.

References

[1] H. Zamani, W. B. Croft, and J. S. Culpepper. "Neural query performance prediction using weak supervision from multiple signals". In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval.* 2018, pp. 105–114.

[2] C. Hauff, L. Azzopardi, and D. Hiemstra. "The combination and evaluation of query performance prediction methods". In: *European Conference on Information Retrieval.* Springer. 2009.

[3] O. Kurland and L. Lee. "PageRank without hyperlinks: Structural reranking using links induced by language models". In: *ACM Transactions on Information Systems (TOIS)* (2010).

[4] O. Kurland and E. Krikon. "The opposite of smoothing: A language model approach to ranking query-specific document clusters". In: *Journal of Artificial Intelligence Research* (2011).

[5] O. Kurland and L. Lee. "Respect my authority! HITS without hyperlinks, utilizing cluster-based language models". In: *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval.* 2006.

[6] E. Krikon and O. Kurland. "A study of the integration of passage-, document-, and cluster-based information for re-ranking search results". In: *Information retrieval* (2011).

[7] E. Krikon, O. Kurland, and M. Bendersky. "Utilizing inter-passage and inter-document similarities for reranking search results". In: *ACM Transactions on Information Systems (TOIS)* (2010).

[8] M. Kusner et al. "From word embeddings to document distances". In: *International conference on machine learning.* PMLR. 2015.

[9] M. Chen. "Efficient vector representation for documents through corruption". In: *arXiv preprint arXiv:1707.02377* (2017).

[10] N. Arabzadeh et al. "Geometric estimation of specificity within embedding spaces". In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management.* 2019.

[11] N. Arabzadeh et al. "Neural Embedding-Based Metrics for Pre-retrieval Query Performance Prediction". In: *European Conference on Information Retrieval.* Springer. 2020.

[12] D. Carmel and E. Yom-Tov. "Estimating the query difficulty for information retrieval". In: *Synthesis Lectures on Information Concepts, Retrieval, and Services* (2010).

[13] D. Metzler and W. B. Croft. "A markov random field model for term dependencies". In: *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval.* 2005.

[14] T. Mikolov et al. "Distributed representations of words and phrases and their compositionality". In: *arXiv preprint arXiv:1310.4546* (2013).