Neural Embedding-Based Metrics for Pre-retrieval Query Performance Prediction

Negar Arabzadeh[1](✉), Fattane Zarrinkalam[1], Jelena Jovanovic[2], and Ebrahim Bagheri[1]

[1] Ryerson University, Toronto, ON, Canada
{narabzad,fzarrinkalam,bagheri}@ryerson.ca
[2] University of Belgrade, Belgrade, Serbia
jelena.jovanovic@fon.bg.ac.rs

Abstract. Query Performance Prediction (QPP) is concerned with estimating the effectiveness of a query within the context of a retrieval model. It allows for operations such as query routing and segmentation, leading to improved retrieval performance. *Pre-retrieval* QPP methods are oblivious to the performance of the retrieval model as they predict query difficulty prior to observing the set of documents retrieved for the query. Since neural embedding-based models are showing wider adoption in the Information Retrieval (IR) community, we propose a set of pre-retrieval QPP metrics based on the properties of *pre-trained* neural embeddings and show that such metrics are more effective for query performance prediction compared to the widely known QPP metrics such as SCQ, PMI and SCS. We report our findings based on Robust04, ClueWeb09 and Gov2 corpora and their associated TREC topics.

Keywords: Query Performance Prediction · Neural embeddings · Specificity

1 Introduction

It is understood that the performance of retrieval models is not always consistent over different queries and corpora and there are some queries that have lower performance, often referred to as *hard* or *difficult* queries [1]. As such, the area of *Query Performance Prediction* is concerned with estimating the performance of a retrieval system for a given query. There is already a well-established body of work that explores query performance prediction through either a *post-retrieval* or a *pre-retrieval* strategy [2]. Methods in post-retrieval measure query difficulty, by analyzing the results obtained from the retrieval system as a response to the query. In contrast, pre-retrieval methods, which are the focus of this work as well, are based on linguistic and statistical features of the query and documents.

While existing work in pre-retrieval query performance has been predominantly focused on defining various statistical measures based on term and corpus-level frequency, the IR community has recently embarked on exploring the impact

J. M. Jose et al. (Eds.): ECIR 2020, LNCS 12036, pp. 78–85, 2020.
https://doi.org/10.1007/978-3-030-45442-5_10

and importance of neural IR techniques [5–7]. There are some recent work that propose to use neural networks for QPP based on a host of signals [8] but to the best of our knowledge, there is only one recent work that specifically utilizes *neural embeddings* of query terms for performing QPP [9]. Neural embeddings maintain interesting *geometric properties* between embedded terms [10] which are manifested by how term vectors are distributed in the embedding space. We explore exploiting the geometric properties of embeddings to define beyond-frequency QPP metrics. Our work *distinguishes* itself from the recent work [9], which proposes to cluster neural embeddings based on their vector similarity to perform QPP, by proposing to not only consider *term similarity* but also take term neighborhood and association into account through a *network representation* of neural embeddings. More specifically, we benefit from term vector associations in the neural embedding space for formalizing *term specificity*, which is correlated with query difficulty [3,4,11].

We base our work on the intuition that a term that has been closely surrounded by several other terms in the embedding space is more likely to be *specific* while a term with a *fewer number* of closely surrounded terms is more likely to be *generic*. We conceptualize the space surrounding a term by using an *ego network* representation where the term of interest serves as the *ego* and is contextualized by a set of *alter* nodes, which are other terms that are similar to it in the embedding space. We apply various measures of node centrality on the ego node to determine the specificity of the term that is being represented by the ego, which would then indicate query difficulty [16]. We have performed experiments based on three widely used TREC corpora, namely Robust04, ClueWeb09 and Gov2 and their corresponding topic sets. Our experiments show that the proposed metrics are effective in QPP using pre-trained neural embeddings.

2 Proposed Approach

This paper is concerned with the design of effective metrics for pre-retrieval QPP based on pre-trained neural embeddings. We focus on distribution of neural embedding vectors in the embedding space to define specificity metrics for QPP. Existing work in the literature [3,12] have already shown that measures of *term specificity* are suitable indicators of query difficulty, i.e., more specific terms are more discriminative and are hence easier to handle when used as queries.

Our work is driven by the *intuition* that more specific terms have a higher likelihood of being surrounded by a larger number of terms compared to generic terms. For instance, as shown in Fig. 1, the set of terms related to the specific term ‘Arsenal’, with an association degree (computed based on cosine similarity of terms’ vector representation) above 0.75, includes terms such as ‘Wenger’, ‘Tottenham’, ‘Everton’, among others, which are also themselves very specific; whereas, the generic term ‘soccer’ has only one closely associated term (association degree above 0.75) and that is ‘football’, which is quite generic itself. While it is not possible to measure frequency information from neural embeddings, it is convenient to identify the set of highly similar terms to a term based on vector similarity. We benefit from this to formalize the notion of an *ego network*

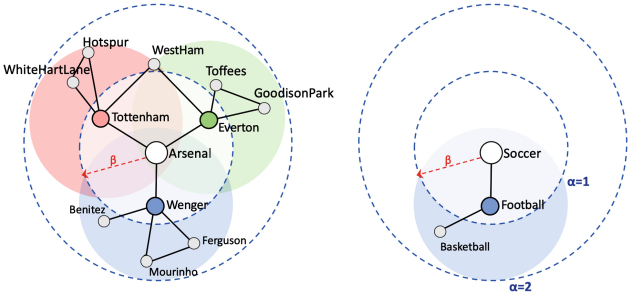

Fig. 1. Schematic of two α-depth β-cut ego networks.

that is based on vector similarities within the embedding space. We benefit from this to formalize our recursive definition of specificity, i.e., the extent to which a term is specific can be determined from the context created by the surrounding highly similar terms within the neural embedding space. In order to formalize *specificity*, we define an *ego network*, as follows:

Definition 1. *Let $\mathcal{P}(t_i, t_j)$ be the degree of similarity between vectors of terms t_i and t_j, V be the complete vocabulary set, and $\mathcal{P}_{\mathcal{M}}(t_i)$ be the highest degree of similarity to t_i from any term in V. We define an $\alpha - depth$ ego network for an ego node t_i in the form of a fully connected graph with a maximum depth α around the ego where the edge weights are $\mathcal{P}(t_k, t_l)$ between any two nodes t_k and t_l. We further refine the $\alpha - depth$ ego network into an $\alpha - depth$ $\beta - cut$ ego network where any edge with a weight less than $\beta \times \mathcal{P}_{\mathcal{M}}(t_i)$ is pruned.*

In simple terms, we propose to build an ego network for a term t_i such that t_i is the ego node and is connected directly to other adjacent terms only if the degree of similarity between the ego and the neighbor is above a discounted rate (β) of the most similar term to the ego. For instance, assuming 'Arsenal' is the ego and $\beta = 0.8$, given that 'Gunners' is the most similar term to the ego with a similarity of 0.854, the immediate neighbors of the ego will consist of all the terms in V that have a similarity above 0.6832 to 'Arsenal'. Furthermore, we allow the ego network to have a depth of α from the ego. For a depth of one ($\alpha = 1$), the ego network will only consist of the ego and its immediate neighbors. For a depth of two ($\alpha = 2$), each node in layer one will become the ego for another sub-ego network with a $\beta - cut$, as explained earlier. Figure 1 shows a schematic of the $\alpha - depth$ $\beta - cut$ *ego network* for the specific term 'Arsenal' and generic term 'soccer'. As seen, in Arsenal's case, the graph is populated with many terms closely related to the ego. In the second layer, the nodes immediately connected to the ego, e.g., 'Wenger', become an ego node for a second layer subgraph, which are in turn connected to their own alters, e.g., 'Mourinho', 'Benitez' and 'Ferguson'. In contrast, the network associated with the generic term 'soccer' is quite sparse with only two additional nodes present when $\alpha = 2$.

Based on the developed ego network, we propose to measure the *specificity* of the ego through the use of *node centrality* metrics [13,16]. Given queries can be

Table 1. Node centrality metrics on the ego network.

Metric	Description
Edge Count (EC)	This metric counts the number of edges in the ego network
Edge Weight Sum (EWS)	This metric calculates the sum of edge weights in the ego network where edge weights are degrees of term association
Inverse Edge Frequency (IEF)	This metric measures the log of the ratio of the number of edges in the network over the number of edges connected to the ego
Degree Centrality (DC)	This metric is the number of links incident upon the ego
Closeness Centrality (CC)	This metric calculates the average length of the shortest path between the ego and all other alters in the network
Betweenness Centrality (BC)	This metric measures the proportion of the shortest paths in the network that go through the ego
Page Rank (PR)	It is based on reciprocity of node importance

composed of more than one term, we adopt the integration approach that uses aggregation functions [14] over the specificity of individual query terms. Table 1 provides an overview of the metrics used in this paper.

3 Experiments

Corpora and Topics: We employed three widely used corpora, namely, Robust04, ClueWeb09, and Gov2. For Robust04, TREC topics 301–450 and 601–650, for Gov2, topics 701–850 and for ClueWeb09, topics 1–200 were used. Topic difficulty was based on Average Precision of each topic computed using QL [15].

Baselines: We adopt the widely used pre-retrieval metrics reported in [2]. The formulation of these metrics is provided in Table 2. As another baseline, we adopt the recent approach by Roy et al. [9] that utilizes embedded word vectors to predict query performance. Their *specificity metric*, known as $P_{clarity}$, is based on the idea that the number of clusters around the neighbourhood of a query term is a potential indicator of its specificity. To apply their approach on our embedding vectors, we have used the implementation provided by the authors.

Neural Embeddings: We used a pre-trained word2vec model based on the Google News corpus (https://goo.gl/wQ8eQ1).

Evaluation: A common approach for measuring the performance of a QPP metric is to use rank correlation metrics to measure the correlation between the list of queries (1) ordered by their difficulty for the retrieval method (ascending order of average precision), and (2) ordered by the QPP metric. Kendall's τ and Pearson's ρ co-efficient are common correlation metrics in this space.

Empirical studies on pre-retrieval QPP metrics have shown that there is no single or set of metrics that outperforms the others on all topics and corpora [2]. Our experiments confirm this. Therefore, to be able to rank the different metrics over a range of topics, we compute the rank of each metric in each topic set and report the rank of the median of each metric over all topics of each document collection. This is specified as *rank* and is reported separately for Kendall's τ and Pearson's ρ. These ranks show how a metric has performed over the different topic sets. Given our metrics are dependent on the α and β parameters, we set them using 5-fold cross validation optimized for Pearson correlation.

Table 2. Baseline metrics. t is a term in query q. d is a document in collection D. D_t is the set of documents with t. $tf(t,D)$ is term frequency of term t in D. $Pr(t|D) = tf(t,D)/|D|$. π_m is the prior probability of the most dominating sense of term t and $P(t|N(\mu_m, \Sigma_m))$ is the posterior probability of term t for the selected cluster.

Metric	Formulation	Ref
IDF	$idf(t) = log(\frac{\vert D\vert}{\vert D_t\vert})$	[2]
VAR	Variance of query term weights $w(t,d)$ in D	[2]
SCQ	$SCQ(t) = (1 + log(tf(t,D))).idf(t)$	[17]
SCS	$SCS(q) = \sum_{t \in q} Pr(t\vert q)log(\frac{Pr(t\vert q)}{Pr(t\vert D)})$	[3]
PMI	$PMI(t_1, t_2) = log\frac{Pr(t_1, t_2\vert D)}{Pr(t_1\vert D)Pr(t_2\vert D)}$	[18]
$P_{clarity}$	$P_{clarity}(t) = \pi_m P(t\vert N(\mu_m, \Sigma_m))$	[9]

Table 3. Results on Robust04. Gray rows are baselines. Bold metrics are the top-3 on Kendall τ (left) and Pearson ρ (right). † indicates statistical significance at alpha = 0.05.

Metric		301-350	351-400	401-450	601-650	Rank	Metric		301-350	351-400	401-450	601-650	Rank
VAR	Avg	0.17	0.36†	0.35†	0.34†	1	VAR	Avg	0.08	0.46†	0.23	0.55†	12
SCQ	Max	0.13	0.42†	0.5†	0.26†	4	SCQ	Max	0.01	0.5†	0.66†	0.35†	4
IDF	Avg	0.22†	0.29†	0.28†	0.32†	5	**IDF**	Avg	0.52†	0.42†	0.43†	0.36†	1
SCS		0.21†	0.24†	0.23†	0.3†	9	SCS		0.43†	0.35†	0.34†	0.42†	7
PMI	Max	0.04	0.18	0.18†	0.2†	12	PMI	Max	0.06	0.12	0.28†	0.08	13
P$_{clarity}$		0.28†	0.2†	0.25†	0.26†	7	P$_{clarity}$		0.34†	0.24†	0.36†	0.38†	9
BC	Max	0.31†	0.2†	0.3†	0.42†	3	BC	Max	0.35†	0.17	0.39†	0.4†	5
IEF	Avg	0.47†	0.36†	0.3†	0.33†	1	**IEF**	Avg	0.41†	0.44†	0.4†	0.43†	1
DC	Max	0.2†	0.29†	0.3†	0.33†	6	DC	Max	0.28†	0.42†	0.44†	0.44†	10
CC	Avg	0.24†	0.29†	0.28†	0.37†	5	CC	Avg	0.26†	0.41†	0.42†	0.45†	5
PR	Max	0.25†	0.4†	0.24†	0.16	8	PR	Max	0.33†	0.47†	0.31†	0.31†	11
EWS	Min	0.17	0.24†	0.29†	0.22†	10	**EWS**	Min	0.34†	0.49†	0.26†	0.46†	3
EC	Min	0.17	0.15	0.27†	0.19	11	EC	Min	0.37†	0.41†	0.26†	0.4†	8

Findings: The results of our experiments are shown in Tables 3, 4 and 5. As shown, our metrics are among the top-3 on both measures on all corpora. On Robust04, two of our metrics, i.e., BC and IEF, are among the top-3 metrics based on Kendall τ. Based on Pearson ρ, IEF and EWS are among the top-3 along with IDF. On Robust04, there is little metric performance consistency on Kendall τ and Pearson ρ. When looking for those metrics that perform well on both measures, IEF and BC are consistent metrics where IEF ranks first on both Kendall τ and Pearson ρ whereas BC ranks third and fifth on these measures, respectively. The other metrics, both baseline metrics and the ones we proposed, have a high performance difference on the two measures. For instance, while the baseline VAR metric ranks first on τ, it ranks twelfth on ρ. On ClueWeb09 and Gov2, unlike Robust04, the top metrics are consistent for Kendall and Pearson where the top-3 metrics include the proposed DC and CC metrics for both measures. On ClueWeb09, these two metrics are accompanied by the BC and

PR metrics for τ and ρ, respectively. However, on Gov2, these metrics are followed by the baseline SCQ metric on τ and our IEF and EWS metrics on ρ. In summary, balancing between the evaluation measures and performance on all topics and corpora, we find our *CC metric* to perform well across the board. It is among the best metrics on Gov2 and ClueWeb09 and has a balanced performance on Robust04. However, CC has a high time complexity of $O(V^3)$. On the other hand, the DC metric performs well on both ClueWeb09 and Gov2 (in the top-3) but less effectiveness on Robust04. The benefit of DC is its low complexity: $O(1)$. Overall, CC is the preferred metric given QPP computations are performed offline. DC can serve as an alternative if computation limitations exist.

Table 4. Results on ClueWeb09. Table format is similar to Table 3.

Metric		1-50	51-100	101-150	150-200	Rank
VAR	Max	0.28†	0.23†	0.27†	0.01	5
SCQ	Max	0.19†	0.25†	0.3†	0.03	4
IDF	Avg	0.24†	0.25†	0.16	0.05	7
SCS		0.23†	0.24†	0.1	0.07	9
PMI	Max	0.15	0.16	0.07	0.03	13
P_clarity		0.3†	0.15	0.1	0.24†	9
BC	Avg	0.22†	0.28†	0.26†	0.22†	3
IEF	Avg	0.17	0.2†	0.17	0.39†	8
DC	Avg	0.22†	0.27†	0.15	0.38†	1
CC	Min	0.22†	0.2†	0.28†	0.36†	2
PR	Max	0.13	0.29†	0.18	0.25†	6
EWS	Max	0.13	0.18	0.26†	0.06	12
EC	Max	0.14	0.16	0.27†	0.05	11

Metric		1-50	51-100	101-150	150-200	Rank
VAR	Max	0.14	0.04	0.42†	0.08	8
SCQ	Max	0.22	0.24	0.33†	0.09	7
IDF	Avg	0.18	0.21	0.27	0.01	10
SCS		0.18	0.19	0.16	0.05	12
PMI	Max	0.2	0.12	0.04	0.04	12
P_clarity		0.29†	0.27	0.18	0.24†	6
BC	Avg	0.28†	0.29†	0.37†	0.28†	4
IEF	Avg	0.29†	0.26	0.17	0.36†	5
DC	Avg	0.25†	0.31†	0.39†	0.34†	3
CC	Min	0.31†	0.17	0.47†	0.33†	1
PR	Max	0.36†	0.31†	0.37†	0.31†	1
EWS	Max	0.07	0.27	0.27	0.14	11
EC	Max	0.06	0.27	0.29	0.14	9

Table 5. Results on Gov2. Table format is similar to Table 3.

Metric		701-750	751-800	801-850	Rank
VAR	Max	0.2†	0.02	0.05	13
SCQ	Max	0.36†	0.29†	0.23†	3
IDF	Avg	0.27†	0.22†	0.14	6
SCS		0.23†	0.19†	0.11	9
PMI	Max	0.28†	0.22†	0.16	6
P_clarity		0.25†	0.24†	0.25†	6
BC	Min	0.18	0.11	0.3†	9
IEF	Max	0.09	0.25†	0.34†	5
DC	Max	0.12	0.36†	0.41†	1
CC	Max	0.11	0.36†	0.41†	1
PR	Min	0.28†	0.19†	0.15	9
EWS	Min	0.2†	0.26†	0.31†	4
EC	Min	0.12	0.17	0.3†	12

Metric		701-750	751-800	801-850	Rank
VAR	Max	0.27	0.06	0.1	13
SCQ	Max	0.53†	0.32†	0.35†	6
IDF	Avg	0.4†	0.25	0.17	10
SCS		0.34†	0.19	0.12	12
PMI	Max	0.44†	0.26	0.22	9
P_clarity		0.33†	0.33†	0.38†	5
BC	Min	0.18	0.28†	0.5†	8
IEF	Max	0.18	0.33†	0.47†	3
DC	Max	0.15	0.41†	0.52†	1
CC	Max	0.14	0.42†	0.51†	1
PR	Min	0.38†	0.32†	0.32†	6
EWS	Min	0.12	0.33†	0.44†	3
EC	Min	0.11	0.24	0.38†	11

4 Concluding Remarks

We have shown that it is possible to devise metrics based on the neural embedding-based representation of terms to perform pre-retrieval QPP. Specifically, we have shown that specificity of a query term, estimated based on an *ego network* representation, can lead to better performance on QPP compared to several baselines such as the one that considers term clusters based on neural embeddings [9].

References

1. Mizzaro, S., Mothe, J.: Why do you think this query is difficult?: A user study on human query prediction. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1073–1076. ACM (2016)
2. Carmel, D., Yom-Tov, E.: Estimating the query difficulty for information retrieval. Synthesis Lectures Inf. Concepts Retrieval Serv. **2**(1), 1–89 (2010)
3. He, B., Ounis, I.: Inferring query performance using pre-retrieval predictors. In: Apostolico, A., Melucci, M. (eds.) SPIRE 2004. LNCS, vol. 3246, pp. 43–54. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-30213-1_5
4. He, J., Larson, M., de Rijke, M.: Using coherence-based measures to predict query difficulty. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 689–694. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-78646-7_80
5. Zuccon, G., Koopman, B., Bruza, P., Azzopardi, L.: Integrating and evaluating neural word embeddings in information retrieval. In: Proceedings of the 20th Australasian Document Computing Symposium, p. 12. ACM (2015)
6. Zhang, L., Zhang, S., Balog, K.: Table2Vec: neural word and entity embeddings for table population and retrieval. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1029–1032. ACM (2019)
7. Mitra, B., Craswell, N.: An introduction to neural information retrieval. Found. Trends Inf. Retrieval **13**(1), 1–126 (2018)
8. Zamani, H., Croft, W.B., Culpepper, J.S.: Neural query performance prediction using weak supervision from multiple signals. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 105–114. ACM (2018)
9. Roy, D., Ganguly, D., Mitra, M., Jones, G.J.: Estimating Gaussian mixture models in the local neighbourhood of embedded word vectors for query performance prediction. Inf. Process. Manage. **56**(3), 1026–1045 (2019)
10. Mimno, D., Thompson, L.: The strange geometry of skip-gram with negative sampling. In: Empirical Methods in Natural Language Processing (2017)
11. Hauff, C., Hiemstra, D., de Jong, F.: A survey of pre-retrieval query performance predictors. In Proceedings of the 17th ACM Conference on Information and Knowledge Management, pp. 1419–1420. ACM (2008)
12. Thomas, P., Scholer, F., Bailey, P., Moffat, A.: Tasks, queries, and rankers in pre-retrieval performance prediction. In: Proceedings of the 22nd Australasian Document Computing Symposium, p. 11. ACM (2017)

13. Segarra, S., Ribeiro, A.: Stability and continuity of centrality measures in weighted graphs. IEEE Trans. Signal Process. **64**(3), 543–555 (2015)
14. Hauff, C., Kelly, D., Azzopardi, L.: A comparison of user and system query performance predictions. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp. 979–988. ACM (2010)
15. Song, F., Croft, W.B.: A general language model for information retrieval. In: Proceedings of the Eighth International Conference on Information and Knowledge Management, pp. 316–321. ACM (1999)
16. Arabzadeh, N., Zarrinkalam, F., Jovanovic, J., Bagheri, E.: Geometric estimation of specificity within embedding spaces. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 2109–2112. ACM (2019)
17. Zhao, Y., Scholer, F., Tsegay, Y.: Effective pre-retrieval query performance prediction using similarity and variability evidence. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 52–64. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-78646-7_8
18. Hauff, C.: Predicting the effectiveness of queries and retrieval systems. In: SIGIR Forum, vol. 44, no. 1, p. 88. ACM (2010)