Novel Query Performance Predictors and their Correlations for Medical Applications

Mohammad Bahrani
Queen Mary University of London
London, UK
m.bahrani@qmul.ac.uk

Thomas Roelleke
Queen Mary University of London
London, UK
t.roelleke@qmul.ac.uk

ABSTRACT

It has been recognized that an obstacle to fully leverage query performance prediction is uncertainty in the effectiveness of the retrieval predictors when applied to different applications of the same task. In this paper we propose novel pre-retrieval predictors that provides formal grounds for the development of a probabilistic framework which serves QPP with respect to various IR models. We explore the influences of different representations of information needs on forecasting the retrieval quality concerning the medical collections. Our study discusses the role of Average Term Frequency, Inverse Document Frequency and the dependency between the query terms in the prediction. We use Dirichlet Multinomial and Natural Harmony assumption to develop new predictors which give rise to the term dependence assumption. Furthermore, we empower the QPP tasks with a position-based TF-IDF measure which potentially enhances the prediction accuracy.

CCS CONCEPTS

• **Information systems → Information retrieval**; *Relevance assessment*; *Retrieval effectiveness*; Test collections;

KEYWORDS

Retrieval predictors, Word burstiness, Medline, TREC, Query Performance Prediction, Evaluation, Correlations, Natural Harmony, Dirichlet multinomial distribution

ACM Reference Format:
Mohammad Bahrani and Thomas Roelleke. 2018. Novel Query Performance Predictors and their Correlations for Medical Applications . In *Proceedings of 2018 (GLARE CIKM).* ACM, New York, NY, USA, 6 pages.

1 INTRODUCTION

Although an Information Retrieval (IR) model might be shown to be effective when applied on a specific medical collection like Medline, there is no way to infer that this model is certainly effective when moved to other collections of the same task such as health-related notifications that are generated from a Body Area Network (BAN). The existence of different representations of clinical information needs is a key parameter that impacts the success rate of IR models [8]. The increasing diversity in the performance of the representations of the information needs led to a new research direction; namely, Query Performance Prediction (QPP) or Query Difficulty Estimation (QDE) [2]. The pre-retrieval predictors such as Average Inverse Document Frequency (AvgIDF), Average Inverse Collection Term Frequency (AvgICTF) and Simplified Clarity Score (SCS) that have been developed are derived from the statistical features of the queries. Interestingly, word burstiness has not been explicitly addressed in the developed predictors.

Within the categorization of medical entities, we often come across various terms for the same concept [1]. Intuitively, if a document starts with a term in relation to a concept and the author intends to repeat the concept, it is more likely that he/she will continue to reuse that specific term. This phenomenon is a type of term dependency which is known as word burstiness [4, 5]. The multinomial probability distribution is a common approach to model the documents but it does not account for word burstiness [5]. Many applications apply the heuristics by topping IR models off with some novel parameters to deliver burstiness identification into the retrieval process. However, these heuristics are not generalizable, and their theoretical explanations are rarely published [13].

In this paper we propose novel pre-retrieval predictors that are based on burstiness identification, position-based term probability and the amount of information carried by the query terms. These predictors build the grounds for the development of a probabilistic framework that predicts the optimal IR models for the given queries against the health-related notifications derived from BANs.

Medline is a commonly used benchmark for searching the biomedical literature. It is maintained by National Library of Medicine (NLM) and as March 2018, it contains more than 24 million references to journal articles [1]. We conducted a set of experiments on the Medline citations and a collection of 25 queries used in 2012, 2011 and 2007 Text Retrieval Conference (TREC) Medical and Genomics tracks in order to capture the correlations between our proposed features and the well-established predictors.

2 RELATED WORK

He and Ounis [7] measured the linear correlations of six pre-retrieval predictors with average precision. Their experimental results showed that the Simplified Clarity Score (SCS) and the Average Inverse Collection Term Frequency (AvgICTF) are the effective predictors. Furthermore, Sondak et al. [15] proposed a QPP framework that gives rise to the effectiveness of the query representation. In [6], He and Ounis studied a set of five pre-retrieval predictors and assessed the linear and non-parametric correlations of the proposed predictors with the query performance. They showed that the effectiveness of a predictor is correlated with the query type. In [10], Mothe and Tanguy investigated the correlation between the query performance and 16 different linguistic features of the query text. Their experimental results showed that syntactic complexity and

[1] https://www.nlm.nih.gov/pubs/factsheets/medline.html

GLARE CIKM, 2018, Turin, Italy

word polysemy are the most significant features. Carmel and Yom-Tov [2] discussed the recent parameters that influence the query difficulty estimation. They compared the correlations between the following pre-retrieval predictors:

AvgIDF. The IDF equation is correlated to one version of Zipf's law which states that if we plot a graph of the log of frequency against the log of rank, the outcome will be a straight line [11].

$$IDF(t) = \log\left(\frac{N}{df(t)}\right). \tag{1}$$

Several pre-retrieval predictors were derived from the IDF quantification such as SumIDF, AvgIDF and MaxIDF. Scholer et al. [14] conducted experiments to predict the query performance based on IDF. They showed that the maximum IDF (MaxIDF) of the query terms provides the highest correlations on TREC web data. Also, the effectiveness of MaxIDF is discussed in [16].

$$SumIDF(q) = \sum_{t \in q} \log\left(\frac{N}{df(t)}\right), \tag{2}$$

$$AvgIDF(q) = \frac{SumIDF(q)}{|q|}, \tag{3}$$

where $df(t)$ denotes the frequency of documents in which the term t is observed.

SCS. Simplified Clarity Score is essentially the relative entropy or Kullback-Leibler (KL) divergence between the query and collection unigram language models [3]. This pre-retrieval predictor has a considerable impact on the performance due to its intrinsic role in the estimation of the query clarity [7].

$$SCS(q) := D_{KL}(q||c) = \sum_{t \in q} p(t|q) \cdot \log_2 \frac{p(t|q)}{p(t|c)}. \tag{4}$$

SCQ. When a query is similar to the corpus, it is easier to retrieve many relevant documents. The Collection Query Similarity measures the similarity between the query and the collection.

$$SCQ(t) = (1 + \log(n(t,c))) \cdot IDF(t), \tag{5}$$

$$MaxSCQ(q) = \max_{t \in q} SCQ(t). \tag{6}$$

AvgPMI. Pointwise Mutual Information is a feature based on the co-occurrence statistics of the query terms. AvgPMI is the average of all PMI scores across possible pairs that can be constructed from the query terms. Accordingly, a high AvgPMI indicates that the query terms are strongly correlated.

3 PROPOSED PRE-RETRIEVAL PREDICTORS

We propose some novel pre-retrieval predictors that leverage word burstiness, position-based term probability and the amount of information carried by the query terms. This section describes AvgTF, PosTF-IDF, DCBackgroundModel and NaturalHarmony as novel predictors which serve the above purposes.

AvgTF. This quantification denotes the Average Term Frequency over *term-elite* documents [9, 12] that are the documents in which the term is observed.

$$AvgTF(t) = \frac{n(t,c)}{df(t)}, \tag{7}$$

where $n(t,c)$ is the wide collection term frequency.

$$SumAvgTF(q) = \sum_{t \in q} AvgTF(t). \tag{8}$$

PosTF − IDF. $TF(t,q)$ is a quantification of the within query term frequency and IDF is the Inverse Document Frequency of term t given the collection. We tune the within query term frequency based on the position of each term in the query, as intuitively the first and last words in a query sequence carry more information.

$$PosTF(t,q) = \begin{cases} n(t,q) + 2, & \text{if } position = 0 \\ n(t,q) + 1, & \text{if } position = n - 1 \\ n(t,q), & \text{otherwise} \end{cases}. \tag{9}$$

$$PosTF - IDF(q) = \sum_{t \in q} PosTF(t) \cdot IDF(t). \tag{10}$$

NaturalHarmony. Based on the independence assumption, the multiple occurrences of event T are assumed to be independent where $p_t^{(n)}$ is the sequence probability to observe n occurrence of event t. Any arbitrary function $f()$ in $p^{f(n)}$ can be employed to represent a form of dependency. In particular, $p_t^{a(n)}$ is the sequence probability where $a(n)$ is the assumption function. [13] described various forms of assumption functions that are based on harmonic sum. Table 1 demonstrates the main harmony assumptions that are derived from a generalized model where α parameter needs to be tuned according to the domain attributes.

$$SumNaturalHarmony(q) = \sum_{t \in q} \left(1 + \frac{1}{2} + .. \frac{1}{n(t,c)}\right). \tag{11}$$

DCBackgroundModel. Cummins et al. [4] brought the term dependence assumption to Language Modelling by using a version of Dirichlet Compound Model which is motivated via the pólya urn process. However, they left the query terms to be treated independently. Equation 12 shows the Corresponding Dirichlet Compound document model and Equation 13 demonstrates the background model that was proposed by them.

$$\alpha_d(t) = \frac{|\overline{d}| \cdot n(t,d)}{|d|}, \tag{12}$$

$$\alpha_c(t) = m_c \cdot \frac{df(t)}{\sum_{i=1}^{n} \left|\overline{d_i}\right|}, \tag{13}$$

$$SumDCBackgroundModel = \sum_{t \in q} \alpha_c(t), \tag{14}$$

where $|\overline{d}|$ is the length of the distinct terms in the document d, $n(t,d)$ is the within document term frequency and $|d|$ is the document length. The estimation of m_c is the requirement for the background model computation. The experiments of *Cummins et al.* [4] suggest to initialize *mc* value to the average document length. They showed

Natural harmony	$1 + \frac{1}{2} + .. \frac{1}{n}$	Harmonic sum
Alpha harmony	$1 + \frac{1}{2^{\alpha}} + .. \frac{1}{n^{\alpha}}$	Generalized harmonic sum
Square root harmony	$1 + \frac{1}{2^{(\frac{1}{2})}} + .. \frac{1}{n^{(\frac{1}{2})}}$	$\alpha = 1/2$; divergent
Square harmony	$1 + \frac{1}{2^2} + .. \frac{1}{n^2}$	$\alpha = 2$; convergent
Gaussian harmony	$2 \cdot \frac{n}{n+1}$	Explains the BM25-TF

Table 1: Main Harmony Assumptions [13].

that within 15 iterations the process will converge. Equation 15 shows the iterative procedure that can be used to estimate *mc*.

$$m_c = \frac{\sum_{i=1}^{n} \left|\overline{d_i}\right|}{\sum_{i=1}^{n} \psi\left(\left|\overline{d_i}\right| + m_c\right) - n \cdot \psi\left(m_c\right)}, \tag{15}$$

where $\psi(x)$ is the digamma function. In this study we sum up the background model weights of the query terms in order to estimate a new feature based on the term dependence assumption.

4 CORRELATIONS BETWEEN THE PREDICTORS

We compared the correlations between the pre-retrieval predictors by experimenting on 25 TREC Medical and Genomics topics and using the *Pearson* coefficient. Not surprisingly, some features correlate with each other and others remain uncorrelated. As can be seen, the results reveal that the highest correlation coefficient is between SumAvgTF and SumNaturalHarmony. However, there are a few cases in the query set that do not follow the pattern which suggests that further experiments can be conducted to estimate the proper predictors for each given query. Although we have not calculated the statistical significance of coefficients, the correlation results may help us to decide which features can potentially be combined to build a performant prediction framework. The correlations between the predictors are shown in table 2. Also, table 3 shows the values of the predictors studied in this paper.

As expected, our experimental results confirm the strong degree of correlation between SumIDF, SumAvgTF and PosTF-IDF which is an indicator of the role of these features as exemplary discriminators. Moreover, our experiments contradict the common assumption which relies on the effectiveness of the query length parameter in the prediction tasks. As an example, table 3 shows that although the "gens are involved in insect segmentation?" query has six terms, the corresponding SumIDF is evidently higher than the "drugs are associated with lysosomal abnormalities in the nervous system" query which consists of ten words.

Predictor	Predictor	Correlation Coefficient
SumAvgTF	SumNaturalHarmony	0.994
SumIDF	PosTF-IDF	0.860
SumIDF	SumAvgTF	0.811
SumIDF	SumNaturalHarmony	0.774
SumNaturalHarmony	SumDCBackgroundModel	0.691
PosTF-IDF	SumNaturalHarmony	0.683
SumAvgTF	SumDCBackgroundModel	0.646
SumIDF	SumDCBackgroundModel	0.404
PosTF-IDF	SumDCBackgroundModel	0.385
PosTF-IDF	MaxSCQ	0.319
SumAvgTF	MaxSCQ	0.237
PosTF-IDF	MaxSCQ	0.188
SumDCBackgroundModel	MaxSCQ	0.138
SumNaturalHarmony	MaxSCQ	00.090
SumIDF	SCS	-0.341
SumAvgTF	SCS	-0.450
PosTF-IDF	SCS	-0.468
SumNaturalHarmony	SCS	-0.500
SumDCBackgroundModel	SCS	-0.591

Table 2: Correlations between the pre-retrieval predictors.

Query	SumIDF	SumAvgTF	PosTF-IDF	SumNaturalHarmony	SumDCBackgroundModel	SCS	MaxSCQ
children with dental caries	6.499	6.401	12.151	37.818	0.395	6.569	29.016
Patients who developed disseminated intravascular coagulation in the hospital	12.060	10.589	18.418	92.349	1.522	-9.598	26.313
patients with inflammatory disorders receiving TNF-inhibitor treatments	8.057	8.133	13.734	68.155	2.723	-6.736	23.048
patients with acute tubular necrosis due to aminoglycoside antibiotics	12.428	11.347	15.091	92.040	3.505	-9.550	28.290
patients who presented to the emergency room with an actual or suspected miscarriage	15.783	11.512	24.141	103.233	3.569	-2.324	28.474
adult inpatients with Alzheimer disease admitted from nursing homes with pressure ulcers	17.077	14.512	21.390	116.352	5.821	-3.987	28.145
patients who have had a carotid endarterectomy	8.635	10.823	13.106	83.484	6.572	-13.810	29.106
patients with hearing loss	4.5406	6.565	9.398	42.295	2.536	-8.117	27.176
hospitalized patients treated for methicillin-resistant Staphylococcus aureus (MRSA) endocarditis	19.188	13.038	20.336	77	3.097	-7.070	32.316
patients diagnosed with localized prostate cancer and treated with robotic surgery	15.278	13.618	25.538	103.780	3.813	-2.432	27.310
women with osteopenia	5.152	3.425	12.388	23.189	0.269	-1.312	28.665
adult patients who received colonoscopies during admission which revealed adenocarcinoma	16.279	14.921	27.222	131.259	8.144	-19.982	26.893
what serum proteins change expression in association with high disease activity in lupus?	14.351	15.811	16.053	139.407	5.460	-6.472	26.421
mutations in the Raf gene are associated with cancer?	6.088	7.405	9.140	59.379	1.564	4.835	24.755
drugs are associated with lysosomal abnormalities in the nervous system?	9.724	8.390	13.211	82.628	2.194	-1.988	26.596

Query	SumIDF	SumAvgTF	PosTF-IDF	SumNaturalHarmony	SumDCBackgroundModel	SCS	MaxSCQ
cell or tissue types express receptor binding sites for vasoactive intestinal peptide (VIP) on their cell surface	17.287	18.647	20.590	152.285	4.273	-10.335	31.107
signs or symptoms of anxiety disorder are related to coronary artery disease	13.134	12.919	18.954	109.430	2.523	-5.316	26.469
toxicities are associated with zoledronic acid	5.106	4.368	8.209	41.816	1.670	-3.593	27.650
gens are involved in insect segmentation?	13.338	4.883	19.168	38.134	0.383	-0.315	28.726
in what diseases of brain development do centrosomal genes play a role?	17.141	12.570	27.287	114.876	2.307	-8.258	27.714
which anaerobic bacterial strains are resistant to Vancomycin?	8.103	8.025	12.918	69.167	3.160	-7.974	29.949
what viral gens affect membrane fusion during hiv infection?	17.021	14.248	29.585	97.979	2.853	-18.305	27.275
what pathways are involved in Ewing's sarcoma?	12.040	6.454	12.040	58.050	0.559	0.577	25.415
what tumor types are found in zebrafish?	9.952	7.4280	12.786	63.456	2.409	-5.767	28.090
proteins make up the murine signal recognition particle	12.838	9.819	19.684	89.250	1.432	-13.388	26.215

Table 3: Values of the pre-retrieval predictors for 25 Medical and Genomics TREC topics.

5 CONCLUSION AND FUTURE WORK

We have introduced a new family of pre-retrieval predictors based on word burstiness, query-position based TF-IDF and Average Term Frequency. We employed a parameter derived from Dirichlet Multinomial Background Model and used the harmony assumption to develop new predictors that capture term dependency and burstiness. We compared the correlations between the proposed features and the well-established pre-retrieval predictors including SumIDF, SCS and MaxSCQ in order to identify the hidden features that may affect the prediction quality. The highest correlation coefficient turned out to be between SumAvgTF and SumNaturalHarmony. As expected, our investigation confirms the strong degree of correlation between SumIDF, SumAvgTF and PosTF-IDF. Surprisingly, all of the predictors remained uncorrelated with SCS which confirms the need of further experiments on SCS. The results will help to learn which predictors are worth being combined in order to increase the prediction accuracy.

Future work will look to evaluate the performance of the proposed predictors on the Medline citations. It could be interesting to detect the effective predictors and subsequently compute their efficiency in some other medical collections. In future work, we also aim to explore the role of Divergence from randomness (DFR) in QPP and discuss the relation between DFR, SCS and Natural Harmony.

REFERENCES

[1] Asma Ben Abacha and Pierre Zweigenbaum. 2011. Automatic extraction of semantic relations between medical entities: a rule based approach. *Journal of biomedical semantics* 2, 5 (2011), S4.

[2] David Carmel and Elad Yom-Tov. 2010. Estimating the query difficulty for information retrieval. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 2, 1 (2010), 1–89.

[3] Steve Cronen-Townsend, Yun Zhou, and W Bruce Croft. 2002. Predicting query performance. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 299–306.

[4] Ronan Cummins, Jiaul H Paik, and Yuanhua Lv. 2015. A Pólya urn document language model for improved information retrieval. *ACM Transactions on Information Systems (TOIS)* 33, 4 (2015), 21.

[5] Charles Elkan. 2006. Clustering documents with an exponential-family approximation of the Dirichlet compound multinomial distribution. In *Proceedings of the 23rd international conference on Machine learning*. ACM, 289–296.

[6] Ben He and Iadh Ounis. 2004. Inferring query performance using pre-retrieval predictors. In *International symposium on string processing and information retrieval*. Springer, 43–54.

[7] Ben He and Iadh Ounis. 2006. Query performance prediction. *Information Systems* 31, 7 (2006), 585–594.

[8] Bevan Koopman, Guido Zuccon, and Peter Bruza. 2017. What makes an effective clinical query and querier? *Journal of the Association for Information Science and Technology* 68, 11 (2017), 2557–2571.

[9] Aldo Lipani, Mihai Lupu, Allan Hanbury, and Akiko Aizawa. 2015. Verboseness fission for BM25 document length normalization. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*. ACM, 385–388.

[10] Josiane Mothe and Ludovic Tanguy. 2005. Linguistic features to predict query difficulty. In *ACM Conference on research and Development in Information Retrieval, SIGIR, Predicting query difficulty-methods and applications workshop*. 7–10.

[11] Stephen Robertson. 2004. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of documentation* 60, 5 (2004), 503–520.

[12] Thomas Roelleke. 2013. Information retrieval models: foundations and relationships. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 5, 3 (2013), 1–163.

[13] Thomas Roelleke, Andreas Kaltenbrunner, and Ricardo Baeza-Yates. 2015. Harmony Assumptions in Information Retrieval and Social Networks. *Comput. J.* 58, 11 (2015), 2982–2999.

[14] Falk Scholer, Hugh E Williams, and Andrew Turpin. 2004. Query association surrogates for web search. *Journal of the American Society for Information Science and Technology* 55, 7 (2004), 637–650.

[15] Mor Sondak, Anna Shtok, and Oren Kurland. 2013. Estimating query representativeness for query-performance prediction. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 853–856.

[16] Ying Zhao, Falk Scholer, and Yohannes Tsegay. 2008. Effective pre-retrieval query performance prediction using similarity and variability evidence. In *European conference on information retrieval*. Springer, 52–64.