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:

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 (AvgDF), 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

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].

\[
IDF(t) = \log\left(\frac{N}{df(t)}\right). 
\]

(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].

\[
SCS(q) := D_{KL}(q||c) = \sum_{t \in q} p(t|q) \cdot \log_{2} \frac{p(t|q)}{p(t|c)} 
\]

\[
SCQ(t) = (1 + \log(n(t,c))) \cdot IDF(t), 
\]

\[
MaxSCQ(q) = \max_{t \in q} SCQ(t). 
\]

### 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.

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

\[
AvgTF(t) = \frac{n(t,c)}{df(t)}, \quad (7)
\]

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

\[
SumAvgTF(q) = \sum_{t \in q} AvgTF(t). \quad (8)
\]

### 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.

\[
PosTF(t,q) = \begin{cases} 
\frac{n(t,q) + 2}{n(t,q) + 1} & \text{if position } = 0 \\
\frac{n(t,q)}{n(t,q) + 1} & \text{if position } = n - 1 \\
\frac{n(t,q)}{n(t,q) + 1} & \text{otherwise}
\end{cases} \quad (9)
\]

### 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 \(p_t^{(n)}\) can be employed to represent a form of dependency. In particular, \(p_t^{(n)}\) 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 \(a\) parameter needs to be tuned according to the domain attributes.

\[
SumNaturalHarmony(q) = \sum_{t \in q} \left(1 + \frac{1}{2} + \ldots + \frac{1}{n(t,c)}\right). \quad (11)
\]

### DCClaimModel

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.

\[
\alpha_y(t) = \frac{|d| \cdot n(t,d)}{|d|}, \quad (12)
\]

\[
\alpha_c(t) = m_c \cdot \frac{df(t)}{\sum_{t \in q} |d|} \quad (13)
\]

\[
SumDCClaimModel = \sum_{t \in q} \alpha_c(t). \quad (14)
\]

where \(|d|\) 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 \(m_c\) value to the average document length. They showed
that within 15 iterations the process will converge. Equation 15 shows the iterative procedure that can be used to estimate \( m_c \).

\[
m_c = \frac{\sum_{i=1}^{n} d_i}{\sum_{i=1}^{n} \psi \left( \frac{d_i}{\sqrt{m_c}} + m_c \right) - n \cdot \psi \left( m_c \right)},
\]

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.
<table>
<thead>
<tr>
<th>Query</th>
<th>SumIDF</th>
<th>SumAvgTF</th>
<th>PostTF-IDF</th>
<th>SumNaturalHarmony</th>
<th>SumDCBackgroundModel</th>
<th>SCS</th>
<th>MaxSCQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>patients with inflammatory disorders receiving TNF-inhibitor treatments</td>
<td>8.057</td>
<td>8.133</td>
<td>13.734</td>
<td>68.155</td>
<td>2.723</td>
<td>-6.736</td>
<td>23.048</td>
</tr>
<tr>
<td>patients who presented to the emergency room with an actual or suspected miscarriage</td>
<td>15.783</td>
<td>11.512</td>
<td>24.141</td>
<td>103.233</td>
<td>3.569</td>
<td>-2.324</td>
<td>28.474</td>
</tr>
<tr>
<td>patients who have had a carotid endarterectomy</td>
<td>8.635</td>
<td>10.823</td>
<td>13.106</td>
<td>83.484</td>
<td>6.572</td>
<td>-13.810</td>
<td>29.106</td>
</tr>
<tr>
<td>patients with hearing loss</td>
<td>4.5406</td>
<td>6.565</td>
<td>9.398</td>
<td>42.295</td>
<td>2.536</td>
<td>-8.117</td>
<td>27.176</td>
</tr>
<tr>
<td>hospitalized patients treated for methicillin-resistant Staphylococcus aureus (MRSA) endocarditis</td>
<td>19.188</td>
<td>13.038</td>
<td>20.336</td>
<td>77</td>
<td>3.097</td>
<td>-7.070</td>
<td>32.316</td>
</tr>
<tr>
<td>patients diagnosed with localized prostate cancer and treated with robotic surgery</td>
<td>15.278</td>
<td>13.618</td>
<td>25.538</td>
<td>103.780</td>
<td>3.813</td>
<td>-2.432</td>
<td>27.310</td>
</tr>
<tr>
<td>women with osteopenia</td>
<td>5.152</td>
<td>3.425</td>
<td>12.388</td>
<td>23.189</td>
<td>0.269</td>
<td>-1.312</td>
<td>28.665</td>
</tr>
<tr>
<td>adult patients who received colonoscopies during admission which revealed adenocarcinoma</td>
<td>16.279</td>
<td>14.921</td>
<td>27.222</td>
<td>131.259</td>
<td>8.144</td>
<td>-19.982</td>
<td>26.893</td>
</tr>
<tr>
<td>drugs are associated with lysosomal abnormalities in the nervous system?</td>
<td>9.724</td>
<td>8.390</td>
<td>13.211</td>
<td>82.628</td>
<td>2.194</td>
<td>-1.988</td>
<td>26.596</td>
</tr>
<tr>
<td>Query</td>
<td>SumIDF</td>
<td>SumAvgTF</td>
<td>PostTF-IDF</td>
<td>SumNaturalHarmony</td>
<td>SumDCBackgroundModel</td>
<td>SCS</td>
<td>MaxSCQ</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>--------</td>
<td>----------</td>
<td>------------</td>
<td>-------------------</td>
<td>----------------------</td>
<td>---------</td>
<td>--------</td>
</tr>
<tr>
<td>cell or tissue types express receptor binding sites for vasoactive intestinal peptide (VIP) on their cell surface</td>
<td>17.287</td>
<td>18.647</td>
<td>20.590</td>
<td>152.285</td>
<td>4.273</td>
<td>-10.335</td>
<td>31.107</td>
</tr>
<tr>
<td>signs or symptoms of anxiety disorder are related to coronary artery disease</td>
<td>13.134</td>
<td>12.919</td>
<td>18.954</td>
<td>109.430</td>
<td>2.523</td>
<td>-5.316</td>
<td>26.469</td>
</tr>
<tr>
<td>toxicities are associated with zoledronic acid</td>
<td>5.106</td>
<td>4.368</td>
<td>8.209</td>
<td>41.816</td>
<td>1.670</td>
<td>-3.593</td>
<td>27.650</td>
</tr>
<tr>
<td>gens are involved in insect segmentation?</td>
<td>13.338</td>
<td>4.883</td>
<td>19.168</td>
<td>38.134</td>
<td>0.383</td>
<td>-0.315</td>
<td>28.726</td>
</tr>
<tr>
<td>in what diseases of brain development do centrosomal genes play a role?</td>
<td>17.141</td>
<td>12.570</td>
<td>27.287</td>
<td>114.876</td>
<td>2.307</td>
<td>-8.258</td>
<td>27.714</td>
</tr>
<tr>
<td>which anaerobic bacterial strains are resistant to Vancomycin?</td>
<td>8.103</td>
<td>8.025</td>
<td>12.918</td>
<td>69.167</td>
<td>3.160</td>
<td>-7.974</td>
<td>29.949</td>
</tr>
<tr>
<td>what pathways are involved in Ewing’s sarcoma?</td>
<td>12.040</td>
<td>6.454</td>
<td>12.040</td>
<td>58.050</td>
<td>0.559</td>
<td>0.577</td>
<td>25.415</td>
</tr>
<tr>
<td>what tumor types are found in zebrafish?</td>
<td>9.952</td>
<td>7.4280</td>
<td>12.786</td>
<td>63.456</td>
<td>2.409</td>
<td>-5.767</td>
<td>28.090</td>
</tr>
</tbody>
</table>

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


