Novel Query Performance Predictors and their Correlations for Medical Applications

Mohammad Bahrani, Thomas Roelleke

Queen Mary University of London

GLARE 2018
Outline

▶ Introduction
▶ Motivation
▶ Methodologies
▶ Experiments and Results
▶ Conclusion and Summary
▶ Future Work
Introduction

- Effectiveness of IR Models on different collections of the same task
Introduction

- Effectiveness of IR Models on different collections of the same task
- Query Performance Prediction (QPP)
Introduction

- Effectiveness of IR Models on different collections of the same task
- Query Performance Prediction (QPP)
- Proposing new retrieval predictors
Introduction

- Effectiveness of IR Models on different collections of the same task
- Query Performance Prediction (QPP)
- Proposing new retrieval predictors
- To build a probabilistic framework for picking the right IR model
Motivation

- Proposing new retrieval predictors.
Motivation

- Proposing new retrieval predictors.
- Apply the retrieval predictors on Medline citations to capture hidden features that may impact QPP.
Motivation

- Proposing new retrieval predictors.
- Apply the retrieval predictors on Medline citations to capture hidden features that may impact QPP.
- To build a probabilistic framework for picking the right IR model.
Methodologies
Standard Pre-Retrieval Predictors

- sumIDF.
Methodologies
Standard Pre-Retrieval Predictors

- sumIDF.
- avgIDF.
Methodologies
Standard Pre-Retrieval Predictors

- sumIDF.
- avgIDF.
- Simplified Clarity Score (KL Divergence).
Methodologies
Standard Pre-Retrieval Predictors

- sumIDF.
- avgIDF.
- Simplified Clarity Score (KL Divergence).
- Collection Query Similarity.
Methodologies

Standard Pre-Retrieval Predictors

- sumIDF.
- avgIDF.
- Simplified Clarity Score (KL Divergence).
- Collection Query Similarity.
- Average Pointwise Mutual Information.
Deficiency
Word businesness and query-position based term probability have not been explicitly addressed.
Methodologies
New Pre-Retrieval Predictors

- **POS-TF.IDF**

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

\[ PosTF - IDF (q) = \sum_{t \in q} PosTF (t) \cdot IDF (t). \]  

(1)  

(2)
Methodologies
New Pre-Retrieval Predictors

- **POS-TF.IDF**

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

  \[
  \text{PosTF} - \text{IDF}(q) = \sum_{t \in q} \text{PosTF}(t) . \text{IDF}(t).
  \]  

- **sumAvgTF**

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

  \[
  \text{SumAvgTF}(q) = \sum_{t \in q} \text{AvgTF}(t).
  \]
Methodologies
New Pre-Retrieval Predictors

- Sum Natural Harmony

\[
\text{SumNaturalHarmony}(q) = \sum_{t \in q} \left( 1 + \frac{1}{2} + \ldots + \frac{1}{n(t, c)} \right).
\] (5)
Methodologies
New Pre-Retrieval Predictors

- Dirichlet Compound Background Model
- Use the sum of DCM background Model
- We need to calculate $mc$

\[
\alpha_d(t) = \frac{|d| \cdot n(t, d)}{|d|},
\]

\[
\alpha_c(t) = m_c \cdot \frac{df_t}{\sum_{i=1}^{n} |d_i|},
\]

\[
SumDCBackgroundModel = \sum_{t \in q} \alpha_c(t).
\]
Experiments and Results

- 25 Medical and Genomics TREC topics
- Applied them on Medline Citations
- Compared the correlation between the retrieval values with the well known predictors:
  - sumIDF
  - SCS
  - maxSCQ
Experiments and Results

Correlations between Predictors

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Predictor</th>
<th>Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>SumAvgTF</td>
<td>SumNaturalHarmony</td>
<td>0.994</td>
</tr>
<tr>
<td>SumIDF</td>
<td>PosTF-IDF</td>
<td>0.860</td>
</tr>
<tr>
<td>SumIDF</td>
<td>SumAvgTF</td>
<td>0.811</td>
</tr>
<tr>
<td>SumIDF</td>
<td>SumNaturalHarmony</td>
<td>0.774</td>
</tr>
<tr>
<td>SumNaturalHarmony</td>
<td>SumDCBackgroundModel</td>
<td>0.691</td>
</tr>
<tr>
<td>PosTF-IDF</td>
<td>SumNaturalHarmony</td>
<td>0.683</td>
</tr>
<tr>
<td>SumAvgTF</td>
<td>SumDCBackgroundModel</td>
<td>0.646</td>
</tr>
<tr>
<td>SumIDF</td>
<td>SumDCBackgroundModel</td>
<td>0.404</td>
</tr>
<tr>
<td>PosTF-IDF</td>
<td>SumDCBackgroundModel</td>
<td>0.385</td>
</tr>
<tr>
<td>PosTF-IDF</td>
<td>MaxSCQ</td>
<td>0.319</td>
</tr>
<tr>
<td>SumAvgTF</td>
<td>MaxSCQ</td>
<td>0.237</td>
</tr>
<tr>
<td>PosTF-IDF</td>
<td>MaxSCQ</td>
<td>0.188</td>
</tr>
<tr>
<td>SumDCBackgroundModel</td>
<td>MaxSCQ</td>
<td>0.138</td>
</tr>
<tr>
<td>SumNaturalHarmony</td>
<td>MaxSCQ</td>
<td>0.00090</td>
</tr>
<tr>
<td>SumIDF</td>
<td>SCS</td>
<td>-0.341</td>
</tr>
<tr>
<td>SumAvgTF</td>
<td>SCS</td>
<td>-0.450</td>
</tr>
<tr>
<td>PosTF-IDF</td>
<td>SCS</td>
<td>-0.468</td>
</tr>
<tr>
<td>SumNaturalHarmony</td>
<td>SCS</td>
<td>-0.500</td>
</tr>
<tr>
<td>SumDCBackgroundModel</td>
<td>SCS</td>
<td>-0.591</td>
</tr>
</tbody>
</table>

Table 2: Correlations between the pre-retrieval predictors.
Experiments and Results
Strong Degree of Correlation

Natural Harmony possess the strongest correlation with other predictors.
- SumAvgTF and sumNaturalHarmony
- sumIDF and sumAvgTF
- sumIDF and sumNaturalHarmony
- sumNaturalHarmony and sumDCBackgroundModel
- Pos-TF-IDF and sumNaturalHarmony
Experiments and Results

Lowest Degree of Correlation

SCS has no correlation with other predictors.

- sumDCBackgroundModel and SCS
- sumNaturalHarmony and SCS
- Pos-TF-IDF and SCS
- sumAvgTF and SCS
- sumIDF and SCS
Summary and Conclusion

- Introduced a new family of pre retrieval predictors.
Summary and Conclusion

- Introduced a new family of pre retrieval predictors.
- Aimed to discover hidden feature of the queries that impact the performance.
Summary and Conclusion

- Introduced a new family of pre retrieval predictors.
- Aimed to discover hidden feature of the queries that impact the performance.
- The highest correlation coefficient was between SumAvgTF and SumNaturalHarmony.
- Surprisingly, all the predictors remained uncorrelated with SCS.
- 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.
Summary and Conclusion

- Introduced a new family of pre retrieval predictors.
- Aimed to discover hidden feature of the queries that impact the performance.
- The highest correlation coefficient was between SumAvgTF and SumNaturalHarmony.
- Surprisingly, all the predictors remained uncorrelated with SCS.

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.
Summary and Conclusion

- Introduced a new family of pre retrieval predictors.
- Aimed to discover hidden feature of the queries that impact the performance.
- The highest correlation coefficient was between SumAvgTF and SumNaturalHarmony.
- Surprisingly, all the predictors remained uncorrelated with SCS.
- Need of further experiments on SCS.
Summary and Conclusion

-Introduced a new family of pre retrieval predictors.
- Aimed to discover hidden feature of the queries that impact the performance.
- The highest correlation coefficient was between SumAvgTF and SumNaturalHarmony.
- Surprisingly, all the predictors remained uncorrelated with SCS.
- 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

- Evaluate the performance of the proposed predictors on the Medline citations.
Future Work

- Evaluate the performance of the proposed predictors on the Medline citations.
- Compute the efficiency of the effective ones in some other medical collections.
Future Work

- Evaluate the performance of the proposed predictors on the Medline citations.
- Compute the efficiency of the effective ones in some other medical collections.
- Aim to explore the role of Divergence From Randomness (DFR) in QPP.
Future Work

- Evaluate the performance of the proposed predictors on the Medline citations.
- Compute the efficiency of the effective ones in some other medical collections.
- Aim to explore the role of Divergence From Randomness (DFR) in QPP.
- Discuss the relation between DFR, SCS and Natural Harmony.
Thank You!