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.

Methodologies

Standard Pre-Retrieval Predictors

Deficiency

Word bustiness and query-position based term probability have not been explicitly addressed.

Methodologies

New Pre-Retrieval Predictors

- POS-TF.IDF

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

$$PosTF-IDF\left(q\right)=\sum_{t\in q}PosTF\left(t\right).IDF\left(t\right). \tag{2}$$

Methodologies

New Pre-Retrieval Predictors

- POS-TF.IDF

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

$$PosTF-IDF\left(q\right)=\sum_{t\in q}PosTF\left(t\right).IDF\left(t\right). \tag{2}$$

- sumAvgTF

$$AvgTF\left(t\right)=\frac{n\left(t,c\right)}{df\left(t\right)}, \tag{3}$$

$$SumAvgTF\left(q\right)=\sum_{t\in q}AvgTF\left(t\right). \tag{4}$$

Methodologies

New Pre-Retrieval Predictors

- Sum Natural Harmony

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

$$SumNaturalHarmony\,(q) = \sum_{t \in q} \left(1 + \frac{1}{2} + .. \frac{1}{n\,(t, c)}\right). \qquad (5)$$

Methodologies

New Pre-Retrieval Predictors

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

$$\alpha_d(t) = \frac{|\overline{d}|.n(t,d)}{|d|}, \tag{6}$$

$$\alpha_c(t) = m_c.\frac{df_t}{\sum_{i=1}^{n}|\overline{d_i}|}, \tag{7}$$

$$SumDCBackgroundModel = \sum_{t\in q}\alpha_c(t). \tag{8}$$

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

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.

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.

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.

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 !