Appendix

ONLINE APPENDIX

Table I shows the QQ metrics selected as features from the paper Predicting Query Quality for Applications of Text Retrieval to Software Engineering Tasks text[1]. Although we only selected nine of the QQ metrics as features, the two approaches we proposed still achieved good results.

TABLE I
Selected Query Quality metrics

Measure	Description	Formula
AvgTFIDF	Average of the term frequency-inverse document frequency (tf-idf)[1] values over all terms in the document	$\frac{1}{\|Q_d\|} \sum_{q \in Q_d} tfidf(q,d)$
MAXTFIDF	Maximum of the term frequency-inverse document frequency (tf-idf) values over all terms in the document	$\max_{q \in Q_d} tfidf(q,d)$
DevTFIDF	The standard deviation of the term frequency-inverse document frequency (tf-idf) values over all terms in the document	$\sqrt{\frac{1}{\|Q_d\|} \sum_{q \in Q_d} (tfidf(q,d) - AvgTFIDF)}$
AvgLogEntropy	Average LogEntropy[2] values over all terms in the document	$\frac{1}{\|Q_d\|} \sum_{q \in Q_d} LogEntropy(q,d)$
MedLogEntropy	Median LogEntropy values over all terms in the document	$\underset{q \in Q_d}{median}\, LogEntropy(q,d)$
DevLogEntropy	The standard deviation of the LogEntropy values over all terms int the document	$\sqrt{\frac{1}{\|Q_d\|} \sum_{q \in Q_d} (LogEntropy(q,d) - AvgLogEntropy)}$
SumSCQ	The sum of the collection-query similarity (SCQ)[3] over all terms in the document	$\sum_{q \in Q_d} (SCQ(q))$
AvgSCQ	The average of the collection-query similarity (SCQ) over all terms in the document	$\frac{1}{\|Q_d\|} \sum_{q \in Q_d} SCQ(q)$
MaxSCQ	The maximum of the collection-query similarity (SCQ) over all terms in the document	$\frac{1}{\|Q_d\|} \max_{q \in Q_d} SCQ(q)$

[1] $tfidf(t,d) = tf(t,d) \cdot \log\left(\frac{|D|}{|D_t|}\right)$

[2] $LogEntropy(t,d) = \log(tf(t,d)+1) \cdot \left(1 + \frac{\sum_{d \in D_t} \frac{tf(t,d)}{tf(t,D)} \cdot \log \frac{tf(t,d)}{tf(t,D)}}{\log(|D|+1)}\right)$

[3] $SCQ(t) = (1 + \log((tf, D))) \cdot \log\left(\frac{|D|}{|D_t|}\right)$

q - a term in the document;

D - the set of documents in the collection;

d - a document in the document collection D;

Q_d - the set of terms in the document d;

D_t - the set of documents containing term t;

$tf(t, D)$ - the frequency of term t in all docs;

$tf(t, d)$ - the frequency of term t in d;

References

[1] Chris Mills, Gabriele Bavota, Sonia Haiduc, Rocco Oliveto, Andrian Marcus, and Andrea De Lucia. 2017. Predicting Query Quality for Applications of Text Retrieval to Software Engineering Tasks.ACM Trans. Softw. Eng. Methodol.26, 1, Article 3 (May 2017), 45 pages.