The 21 Pre-Retrieval Query Measures

Property	Measure	Description	Formula
Specificity	*AvgIDF*	Average of the Inverse Document Frequency (*idf*) values over all query terms	$\frac{1}{\lvert Q\rvert}\sum_{q\in Q} idf(q)$
	MaxIDF	Maximum of the Inverse Document Frequency (*idf*) values over all query terms	$max_{q\in Q}(idf(q))$
	DevIDF	The standard deviation of the Inverse Document Frequency (*idf*) values over all query terms	$\sqrt{\frac{1}{\lvert Q\rvert}\sum_{q\in Q}(idf(q)-avgIDF)}$
	AvgICTF	Average Inverse Collection Term Frequency (*ictf*) values over all query terms	$\frac{1}{\lvert Q\rvert}\sum_{q\in Q} ictf(q)$
	MaxICTF	Maximum Inverse Collection Term Frequency (*ictf*) values over all query terms	$max_{q\in Q}(ictf(q))$
	DevICTF	The standard deviation of the Inverse Collection Term Frequency (*ictf*) values over all query terms	$\sqrt{\frac{1}{\lvert Q\rvert}\sum_{q\in Q}(ictf(q)-avgICTF)}$
	AvgEntropy	Average *entropy* values over all query terms	$\frac{1}{\lvert Q\rvert}\sum_{q\in Q} entropy(q)$
	MedEntropy	Median *entropy* values over all query terms	$median_{q\in Q}\ (entropy(q))$
	MaxEntropy	Maximum *entropy* values over all query terms	$max_{q\in Q}\ (entropy(q))$
	DevEntropy	The standard deviation of the *entropy* values over all query terms	$\sqrt{\frac{1}{\lvert Q\rvert}\sum_{q\in Q}(entropy(q)-avgEntropy)}$
	QS	Query Scope – the percentage of documents in the collection containing at least one of the query terms	$\frac{\lvert \cup_{q\in Q} D_q \rvert}{\lvert D\rvert}$
	SCS	Simplified Clarity Score – the Kullback-Leiber divergence of the query language model from the collection language model	$\sum_{q\in Q} p_q(Q)\cdot\log(\frac{p_q(Q)}{p_q(D)})$
Coherency	*AvgVAR*	Average of the variances of the query term weights over the documents containing the query term (*VAR*), over all query terms	$\frac{1}{\lvert Q\rvert}\sum_{q\in Q} VAR(q)$
	MaxVAR	Maximum of the variances of the query term weights over the documents containing the query term (*VAR*), over all query terms	$max_{q\in Q}(VAR(q))$
	SumVAR	Sum of the variances of the query term weights over the documents containing the query term (*VAR*), over all query terms	$\sum_{q\in Q} VAR(q)$
	CS	Coherence Score – the average of the pairwise similarity between all pairs of documents containing one of the query terms (*cs*) among all	$\frac{1}{\lvert Q\rvert}\sum_{q\in Q} cs(q)$
Similarity	*AvgSCQ*	The average of the collection-query similarity (*SCQ*) over all query terms	$\frac{1}{\lvert Q\rvert}\sum_{q\in Q} SCQ(q)$
	MaxSCQ	The maximum of the collection-query similarity (*SCQ*) over all query terms	$max_{q\in Q}(SCQ(q))$
	SumSCQ	The sum of the collection-query similarity (*SCQ*) over all query terms	$\sum_{q\in Q} SCQ(q)$
Term relatedness	*AvgPMI*	Average Pointwise Mutual Information (*PMI*) over all pairs of terms in the query	$\frac{2\,(\lvert Q\rvert-2)!}{(\lvert Q\rvert)!}\sum_{q_1,q_2\in Q} PMI(q_1,q_2)$
	MaxPMI	Maximum Pointwise Mutual Information (*PMI*) over all pairs of terms in the query	$max_{q_1,q_2\in Q}(PMI(q_1,q_2))$

$idf(t)=\log(\frac{\lvert D\rvert}{\lvert D_t\rvert})$	$p_t(d)=\frac{tf(t,d)}{\lvert d\rvert}$	$w(t,d)=\frac{1}{\lvert d\rvert}\log(1+tf(t,d))\cdot idf(t)$	$entropy(t)=\sum_{d\in D_t} p_t(d)\cdot \log_{\lvert D\rvert} p_t(d)$
$ictf(t)=\log(\frac{\lvert D\rvert}{tf(t,D)})$	$p_t(D)=\frac{tf(t,D)}{\lvert D\rvert}$	$SCQ(t)=(1+\log(ctf(t,D))\cdot idf(t)$	$PMI(t_1,t_2)=log\frac{p_{t_1,t_2}(D)}{p_{t_1}(D)\cdot p_{t_2}(D)}$
$\bar{w}_t=\frac{1}{\lvert D_t\rvert}\sum_{d\in D_t} w(t,d)$	$p_t(Q)=\frac{tf(t,Q)}{\lvert Q\rvert}$	$cs(t)=\frac{\sum_{(d_i,d_j)\in D_t} sim(d_i,d_j)}{\lvert D_t\rvert\cdot(\lvert D_t\rvert-1)}$	$VAR(t)=\sqrt{\frac{\sum_{d\in D_t}(w(t,d)-\bar{w}_t)^2}{df(t)}}$

Q –the set of query terms; q – a term in the query; D – the set of documents in the collection; D_t –the set of documents containing term t

d – a document in the document collection D; $tf(t,D)$ – the frequency of term t in all docs; $tf(t,d)$ – the frequency of term t in d

$tf(t,Q)$ – the frequency of term t in the query; $sim(d_i,d_j)$ – the cosine similarity between the vector-space representations of d_i and d_j