Handout of the paper "Exploring the Similarity Space" by Zobel and Moffat

Related Terms

- the number N of documents;
- the number n of distinct terms used in the collection;
- for each term t and each document d containing t, the frequency $f_{d,t}$ of t in d;
- for each term t, the total number F_t of occurrences of t in the collection;
- the number ft of documents containing term t;
- for each term t, the frequency $f_{q,t}$ of t in query q;
- for each document d, the value $f_d = |d|$, the number of term occurrences in d;
- for each document d, f^m_d , the largest $f_{d,t}$ of any term in d; and
- f^m, the largest f_t in the collection.

Documents and terms are then gathered into sets that restrict the domain of the operations used to combine the statistics into similarity values. We denote these various sets as:

- the set *D* of documents;
- for each term *t*, the set D_t of documents containing *t*;
- the set *T* of distinct terms in the database; and
- the set T_d of distinct terms in document *d*, and similarly T_q for queries, and $T_{q;d} = T_q \setminus T_d$.

	Description	Formulation
A	Inner product.	$S_{q,d} = \sum_{t \in \mathcal{T}_{q,d}} (w_{q,t} \cdot w_{d,t})$
B	Cosine measure.	$S_{q,d} = \frac{\sum_{t \in \mathcal{T}_{q,d}} (w_{q,t} \cdot w_{d,t})}{W_q \cdot W_d}$
C	Simple probabilistic measure. The variable C is a tuning constant, set to 0 in this context [Frakes and Baeza-Yates 1992, p. 369].	$S_{q,d} = \sum_{t \in \mathcal{T}_{q,d}} (C + w_t)$
D	More sophisticated probabilistic measure. Variable C is again a tuning constant set to 0.	$S_{q,d} = \sum_{t \in \mathcal{T}_{q,d}} (C + w_t) \cdot r_{d,t}$
E	Alternative inner product.	$S_{q,d} = \sum_{t \in \mathcal{T}_{q,d}} \frac{w_{d,t}}{W_d}$
F	Dice formulation. (Ozkarahan [1986, p. 496] and Salton and McGill [1983, pp. 202–3] use $W_x = \sum_{t \in \mathcal{T}_x} w_{x,t}$ rather than W_x^2, for Dice, Jaccard, and overlap.)	$S_{q,d} = \frac{2 \sum_{t \in \mathcal{T}_{q,d}} (w_{q,t} \cdot w_{d,t})}{W_q^2 + W_d^2}$
G	Jaccard formulation.	$S_{q,d} = \frac{\sum_{t \in \mathcal{T}_{q,d}} (w_{q,t} \cdot w_{d,t})}{W_q^2 + W_d^2 - \sum_{t \in \mathcal{T}_{q,d}} (w_{q,t} \cdot w_{d,t})}$
H	Overlap formulation.	$S_{q,d} = \frac{\sum_{t \in \mathcal{T}_{q,d}} (w_{q,t} \cdot w_{d,t})}{\min(W_q^2, W_d^2)}$

Table 1: Combining functions $S_{q,d}$