解説

tfは term frequncyの略で、ある文章中に出現する用語tの頻度の事である。

単純頻度ならtfは単語が文章で出てきた回数となる。

相対頻度なら 単語の単純頻度/文の用語数(dl : 文章長と呼ばれる)

二値でtfを表現 tf=1（出現）/0（not 出現）

logを使ってtfを表現 tf ={ 1+log(単純回数) | 0}

idf 逆文書頻度の略。文章にどれくらい出現するか たくさん出現すればidfは下げる。

idf = log(N/df) N:全文書 df:出現する文書数