Frekuensi Jangka
Frekuensi Dokumen Terbalik

Ridwan Mahenra, S.Kom., M.C.S(AI)

Fakultas Teknik dan Ilmu Komputer
Universitas Teknokrat Indonesia

TF-IDF

Dalam temu kembali informasi, tf-idf, TF*IDF, atau TFIDF (singkatan dari bahasa Inggris: term frekuensi-invers dokumen frekuensi, bahasa Indonesia: term frekuensi-invers dokumen frekuensi) adalah ukuran statistik yang menggambarkan pentingnya suatu istilah terhadap suatu dokumen dalam suatu koleksi atau korpus. Hal ini sering digunakan sebagai faktor pembobotan dalam pengambilan informasi, penambangan teks, dan pemodelan pengguna. Nilai tf-idf meningkat sebanding dengan jumlah kemunculan term in dan bergantung pada jumlah dokumen dalam korpus yang memiliki term tersebut.

Frekuensi Jangka

Frekuensi suku, tf(t, d), *t* adalah frekuensi suku

$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

dimana **ft,d** adalah pencacahan mentah istilah-istilah dalam dokumen, yaitu banyaknya kemunculan istilah ***t*** dalam dokumen d. Semakin sering suatu term muncul maka semakin besar nilai tfnya. Ada beberapa cara untuk mendefinisikan frekuensi istilah.

Variasi tertimbang frekuensi suku (tf)

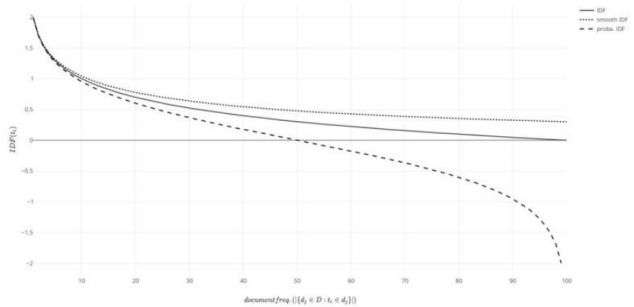
skema	pembobotan tf
Biner	$0, 1$
enumerasi mentah	$f_{t,d}$
frekuensi istilah	$\frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$
normalisasi log	$\log(1 + f_{t,d})$
normalisasi ganda 0,5	$0{,}5 + 0{,}5 \times \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$
normalisasi ganda K	$K + (1 - K)\frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$

Frekuensi Dokumen Terbalik

$$\mathrm{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

Invers frekuensi dokumen, ***idf(t,D),*** adalah ukuran informasi yang disediakan oleh istilah t, yaitu seberapa sering atau jarang suatu istilah muncul di seluruh dokumen. Semakin jarang suatu istilah di antara dokumen, semakin besar nilai idfnya. Nilainya merupakan logaritma kebalikan dari banyaknya dokumen yang mempunyai term **t** dibagi dengan jumlah seluruh dokumen (N), dimana

himpunan **{d** ÿ **D: t** ÿ **d}** adalah himpunan dokumen **d** di **D** yang mempunyai istilah t.



Grafik berbagai fungsi inversi frekuensi dokumen: standar, mulus, probabilistik

Ragam bobot inversi frekuensi dokumen (idf)

skema	pembobotan idf
dasar satu	1
inversi frekuensi dokumen	$\log \frac{N}{n_t} = -\log \frac{n_t}{N}$
mulus	$\log\left(\frac{N}{1+n_t}\right)+1$
maks	$\log\left(\frac{\max_{\{t' \in d\}} n_{t'}}{1+n_t}\right)$
probabilistik	$\log \frac{N-n_t}{n_t}$
Keterangan: *nt* = |{d ÿ D: *t* ÿ d}|	

Istilah frekuensi-inversi frekuensi dokumen

Istilah inversi frekuensi nilai frekuensi dokumen (tf-idf) dapat dihitung dengan cara

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \times \mathrm{idf}(t, D).$$

Nilai ini akan besar jika term sering muncul (tf besar), namun hanya di beberapa dokumen (idf besar atau df kecil). Nilai ini biasanya membuang istilah-istilah umum.

skema	bobot jangka dokumen (d)	bobot istilah kueri (q)
1	$f_{t,d} \cdot \log \frac{N}{n_t}$	$\left(0{,}5 + 0{,}5 \times \frac{f_{t,q}}{\max_t f_{t,q}}\right) \times \log \frac{N}{n_t}$
2	$\log(1 + f_{t,d})$	$\log\left(1 + \frac{N}{n_t}\right)$
3	$(1 + \log f_{t,d}) \times \log \frac{N}{n_t}$	$(1 + \log f_{t,q}) \times \log \frac{N}{n_t}$

Peran IDF

Nilai IDF diperkenalkan sebagai "istilah kekhasan" oleh Karen Spärck Jones dalam sebuah makalah tahun 1972. Meskipun itu berfungsi dengan baik sebuah heuristik, landasan teoretisnya telah menjadi masalah selama setidaknya tiga puluh tahun. Para peneliti sedang mencoba menjelaskannya secara informasi-teoretis.

Penjelasan Spärck Jones sendiri sebenarnya tidak banyak mengajukan teori selain kaitannya dengan hukum Zipf.
Beberapa upaya telah dilakukan untuk menempatkan idf dalam bidang probabilistik dengan memperkirakan probabilitas bahwa a dokumen d mempunyai term t sebagai frekuensi relatif dokumen tersebut,

$$P(t|D) = \frac{|\{d \in D : t \in d\}|}{N},$$

jadi kita bisa mendefinisikan idf sebagai berikut.

$$\begin{aligned} \text{idf} &= -\log P(t|D) \\ &= \log \frac{1}{P(t|D)} \\ &= \log \frac{N}{|\{d \in D : t \in d\}|} \end{aligned}$$

Dengan kata lain, inversi frekuensi dokumen adalah logaritma dari "inversi" frekuensi relatif dokumen.

Penafsiran probabilistik ini menggunakan bentuk yang sama dengan isi informasi.
Namun, menerapkan pandangan teori informasi pada masalah dalam pengambilan informasi menyebabkan masalah ketika mencoba menentukan ruang sampel untuk distribusi probabilitas: tidak hanya dokumen yang harus dipertimbangkan, tetapi juga pertanyaan dan istilah.

Contoh tf-idf

Misalkan terdapat tabel jumlah suku dalam suatu korpus yang berisi dua dokumen seperti tabel di samping.

Untuk menghitung tf-idf dari term **“ini”** dapat dilakukan langkah-langkah sebagai berikut.

Dalam bentuk frekuensi mentahnya, tf hanyalah frekuensi istilah **"ini"** di setiap dokumen. Di setiap dokumen, istilah **“ini”** muncul satu kali.

Namun, karena dokumen 2 memiliki lebih banyak kata, frekuensi relatifnya lebih kecil.

Dokumen 1

Jangka Waktu Total

ini	1
adalah	1
sebuah	2
sampel	1

Dokumen 2

Jangka Waktu Total

ini	1
adalah	1
contoh	3
lainnya	2

$$\text{tf}(''\text{ini}'', d_1) = \frac{1}{5} = 0{,}2$$
$$\text{tf}(''\text{ini}'', d_2) = \frac{1}{7} \approx 0{,}14$$

Nilai idfnya tetap per korpus dan bergantung pada jumlah dokumen yang memiliki istilah **"ini".** Dalam hal ini, kami memiliki korpus yang semua dokumennya memiliki istilah **"ini".**

$$\text{idf}(''\text{ini}'', D) = \log\left(\frac{2}{2}\right) = 0$$

Jadi, nilai tf-idf istilah ini adalah nol yang berarti istilah ini tidak terlalu bermakna seperti yang terlihat di keseluruhan dokumen.

$$\text{tfidf}(''\text{ini}'', d_1, D) = 0{,}2 \times 0 = 0$$
$$\text{tfidf}(''\text{ini}'', d_2, D) = 0{,}14 \times 0 = 0$$