Tìm kiếm thông tin và trình diễn thông tin

Bài 5. Đánh giá kết quả tìm kiếm

Nội dung

Tóm tắt nội dung liên quan

1. P, R và độ đo F

2. P/R, nội suy, AP

3. Bộ dữ liệu kiểm thử

Ví dụ phù hợp Boolean

Truy vấn: ((*văn bản* ∨ *thông tin*) ∧ *tìm kiếm* ∧ ¬*lý thuyết*)

Văn bản:

1. *"Tìm kiếm thông tin*
2. *"Lý thuyết thông tin"*
3. *"Tìm kiếm thông tin hiện đại: lý thuyết và thực hành"*
4. *"Phương pháp nén văn bản"*

Mô hình không gian vec-tơ: hệ ký hiệu SMART

Term frequency		Document frequency		Normalization	
n (natural)	$\mathrm{tf}_{t,d}$	n (no)	1	n (none)	1
l (logarithm)	$1 + \log(\mathrm{tf}_{t,d})$	t (idf)	$\log \frac{N}{\mathrm{df}_t}$	c (cosine)	$\frac{1}{\sqrt{w_1^2+w_2^2+\ldots+w_M^2}}$
a (augmented)	$0.5 + \frac{0.5 \times \mathrm{tf}_{t,d}}{\max_t(\mathrm{tf}_{t,d})}$	p (prob idf)	$\max\{0, \log \frac{N-\mathrm{df}_t}{\mathrm{df}_t}\}$	u (pivoted unique)	$1/u$
b (boolean)	$\begin{cases} 1 & \text{if } \mathrm{tf}_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$			b (byte size)	$1/CharLength^{\alpha}$, $\alpha < 1$
L (log ave)	$\frac{1+\log(\mathrm{tf}_{t,d})}{1+\log(\mathrm{ave}_{t \in d}(\mathrm{tf}_{t,d}))}$				

Mô hình BIM

$$RSV(d,q) = \sum_{x_i = q_i = 1} c_i$$

$$c_t = \log \frac{(s+0.5)(N-S-n+s+0.5)}{(n-s+0.5)(S-s+0.5)}$$

Mô hình Okapi BM25

$$RSV_d = \sum_{t \in q} \left[\log \left[\frac{(s + 1/2)/(S - s + 1/2)}{(n - s + 1/2)/(N - n - S + s + 1/2)} \right] \times \right.$$

$$\left. \times \frac{(k_1 + 1) tf_{t,d}}{k_1 \left((1 - b) + b \times (L_d / L_{ave}) \right) + tf_{t,d}} \times \frac{(k_3 + 1) tf_{t,q}}{k_3 + tf_{t,q}} \right]$$

Mô hình ngôn ngữ:

Đa thức, đơn từ, làm mịn tuyến tính

$$RSV(q,d) = \prod_{t \in q} \left(\lambda \frac{tf_{t,d}}{L_d} + (1-\lambda) \frac{cf_{t,C}}{L_C} \right)$$

P, R, F

Precision = #(văn bản phù hợp trả về)/#(văn bản trả về)

Recall = #(văn bản phù hợp trả về)/#(văn bản phù hợp)

Ký hiệu P: độ chính xác; R: độ đầy đủ

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}}$$

$$\beta^2 = \frac{1-\alpha}{\alpha}$$

$$F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

P/R, nội suy, AP

P@i = #(văn bản phù hợp trong i kết quả đầu tiên)/i

R@i = #(văn bản phù hợp trong i kết quả đầu tiên)/
#(số văn bản phù hợp trong bộ dữ liệu)

$$p_{\text{inter}}(r) = \max_{r^* \geq r} p(r^*)$$

$$MAP = \frac{1}{|Q|} \cdot \sum \left(\frac{1}{R_q} \cdot \sum P@K_i\right)$$

Bộ dữ liệu kiểm thử

TABLE 4.3 Common Test Corpora

Collection	*NDocs*	*NQrys*	*Size (MB)*	*Term/Doc*	*Q-D RelAss*
ADI	82	35			
AIT	2109	14	2	400	>10,000
CACM	3204	64	2	24.5	
CISI	1460	112	2	46.5	
Cranfield	1400	225	2	53.1	
LISA	5872	35	3		
Medline	1033	30	1		
NPL	11,429	93	3		
OSHMED	34,8566	106	400	250	16,140
Reuters	21,578	672	28	131	
TREC	740,000	200	2000	89-3543	» 100,000

		Đánh giá 2		
		Yes	No	Total
Đánh giá 1	Yes	300	20	320
	No	10	70	80
	Total	310	90	400

Theo dõi tỉ lệ số lần thống nhất của kết quả

$P(A) = (300 + 70)/400 = 370/400 = 0.925$

Giá trị biên tổng hợp

$P(\textit{không phù hợp}) = (80 + 90)/(400 + 400) = 170/800 = 0.2125$

$P(\textit{phù hợp}) = (320 + 310)/(400 + 400) = 630/800 = 0.7878$

Giá trị xác suất của sự thống nhất ngẫu nhiên $P(E)$ =

$P(\textit{không phù hợp})^2 + P(\text{phù hợp})^2 = 0.21252^2 + 0.78782^2 = 0.665$

Chỉ số kappa $\kappa = (P(A) - P(E))/(1 - P(E)) =$

$(0.925 - 0.665)/(1 - 0.665) = 0.776$ (trong khoảng được chấp nhận)

Câu hỏi & Thảo luận?