Nikitha M S
4NI17IS051
ISE 8 'A'
IS0328 - IRS

1 logical view of document

Documents in a collection are represented through set of index terms or key words Such keywords might be extracted directly from the text of the document. No matter if representative keywords are derived automatically or generated by a specialist they provide a logical view of the document

The user of retrieval system has to translate his information need into a query in the language provided by the system. We make a clear distinction between the different task the user of the retrieval system might be engaged in the task might be of 2 different type Information or data retrieval and browsing : Data retrieval system a query expression is used to convey the constraints that must be satisfied by objects in the answer set. Browsing in language of world wide web pulling action where the user request the information in an interactive manner, or by using software agent which push information towards the user.

4 Precision : It is defined as the ratio or fraction of the retrieved documents which is relevant

|A| no of doc in the set

$$\text{Precision} = \frac{|R_a|}{|A|}$$

|Ra| - No of doc in intersection of set R & A

Recall : It is defined as the fraction of the relevant documents which has been retrieved

|R| - no of doc in the set

$$\text{Recall} = \frac{|R_a|}{|R|}$$

$$A = \{ d_{28}, d_{39}, d_1, d_{32}, d_{98}, d_{29}, d_{81}, d_{93} \}$$

$$R = \{ d_9, d_{18}, d_{32}, d_{40}, d_{29}, d_{39}, d_{92}, d_{28} \}$$

$$|A| = 8 \qquad |R| = 8$$

$$\{R_a\} = \{ d_{28}, d_{29}, d_{32}, d_{39} \}$$

$$|R_a| = 4.$$

1 of 3

Nikitha M S
4NI17IS051
ISE 8 'A'
IS0S28 - IRS

$$\text{Precision} = \frac{4}{8} = \frac{1}{2}.$$

$$\text{Recall} = \frac{|Ra|}{|R|} = \frac{\cancel{4}}{8_2} = \frac{1}{2}.$$

5 R-precision : To generate a single value summary of the ranking by computing precision at R^{th} position in the ranking, where R is the total number of relevant documents for a current query. The R-precision measure is useful parameter for observing the behavior of an algorithm for each individual query in an experiment

Answer set $_{Alg\ A}$ → $\{d_9, d_{39}, d_{38}, d_{26}, d_1\}$

Answer set $_{Alg\ B}$ → $\{d_{28}, d_{18}, d_{40}, d_{27}, d_9, d_{20}\}$

R → $\{d_9, d_{18}, d_{32}, d_{40}, d_{29}, d_{39}, d_{92}, d_{28}\}$.

$|R| = 8$

$$RP_A = \frac{2}{8} = \frac{1}{4} = 0.25$$

$$RP_B = \frac{4}{8} = \frac{1}{2} = 0.5$$

2 D_1 = Cat and rat are animals

D_2 = Animals chased their Pray

Q = Cat chased rat.

freq (i, j).

N - Total no. of doc

n_i - no of doc which contain the Keyword K_i

$idf = \log\left(\frac{N}{n_i}\right)$

$w_{i,j}$: $freq(i,j) * idf$

$$sim(d_j, q) = \frac{\vec{d_j} \cdot \vec{q}}{|\vec{d_j}| \cdot |\vec{q}|}$$

$$= \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}}$$

2 of 3

Nikitha MS
4NI17IS051
ISE 8 'A'
IS0328 - IRS

3 document similarity $\text{sim}(D, Q)$

~~$$\text{sim}(d_j, q) = \frac{\vec{d_j} \cdot \vec{q}}{|\vec{d_j}| \cdot |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}}$$~~

$$\text{Sim}(Q, D) = \prod \frac{P(t_i/R)}{P(t_i/\bar{R})} \frac{P(R)}{P(\bar{R})}$$

Documents	Revelence	t_1	t_2
D_1	R	0	1
D_2	NR	1	0
D_3	NR	1	1
D_4	R	1	0
D_5	NR	1	1

Relevant → $\overset{t_2}{D_1}$ $\overset{t_1}{D_4}$

non relevant → D_2 D_3 D_5

$D_2 \downarrow t_1$, $D_3 \downarrow t_1, t_2$, $D_5 \downarrow t_1, t_2$

$$\text{Sim}(Q, D_2) = \frac{1/2}{2/3} \cdot \frac{2/5}{3/5} = 0.5$$

$$\text{Sim}(Q, D_3) = \frac{1/2}{2/3} \cdot \frac{2/\cancel{5}}{3/\cancel{5}} \cdot \frac{\cancel{2}/\cancel{2}}{1/3} \cdot \frac{2/\cancel{5}}{3/\cancel{5}}$$

$$= \frac{\cancel{3}}{\cancel{4}} \times \frac{\cancel{2}}{\cancel{3}} \times \cancel{3} \times \frac{\cancel{2}}{\cancel{3}}$$

$$= 1$$

3 of 3