

Prova

AND LOGICO

1 1 0 1 0 0 AND

1 1 0 1 1 1 AND

1 0 1 1 1

1 0 0 1 0 0

① LAVAGNA DEL 20/10/20

DATI UNA COLLEZIONE DI DOCUMENTI E LE RELATIVE OCCORRENZE DEI TERMINI

$$d_1 = (T_1: 7, T_2: 3, T_5: 1)$$

$$d_2 = (T_1: 4, T_5: 3)$$

$$d_3 = (T_1: 2, T_3: 2, T_4: 1)$$

$$d_4 = (T_3: 2)$$

A) COSTRUIRE L'INDICE INVERTITO DELLA COLEZIONE

... DOLLA SELECTIONS

B) CALCOLARE IL RANKING DEI DOCUMENTI RISPETTO ALLA QUERY

$$q \geq (T1:1 \text{ AND } TS:2)$$

A) COSTRUZIONE INDICE INVERTITO

$T_1^0: D_1, D_2, D_3, D_4$

$T_2^0: D_1$

$T_3^0: D_3, D_4$

$T_4^0: D_3$

$T_5^0: D_1, D_2$

B) PER AISOLVERE Q SISTO QUESTA PROCEDURA

- RECUPERO LA POSTING LIST DI T1

$T_1^0: D_1, D_2, D_3, D_4$

- RECUPERO LA POSTING LIST DI TS

$T_5 \cap D_1, D_2$

- FAÇCIO L'INTERSEZIONE

$$T_1 \cap T_5 = D_1, D_2$$

- CALCOLO LA SOMIGLIANZA DEL PRODOTTO

INTERNO TRA¹

$$\vec{D}_1 = (7 \ 5 \ 0 \ 0 \ 1)$$

$$\vec{D}_2 = (4 \ 0 \ 0 \ 0 \ 3)$$

VETTORI SECONDO PAROLE
ALL'INTERNO DEI DOCUMENTI.

GUARDA TRACCIA

$$\vec{Q} = (1 \ 0 \ 0 \ 0 \ 2) \rightarrow \text{VETTORE QUERY}$$

P.S. HO 5 PAROLE T_1, T_2, T_3, T_4, T_5 . METTO Ø IN

CORRISPONDENZA DELLE PAROLE NON PRESENTI NELLA QUERY Ø NEI DOCUMENTI

$$1 \text{ SIM}(Q, D_1) = 7 \cdot 1 + 5 \cdot 0 + 0 \cdot 0 + 0 \cdot 0 + 1 \cdot 2 = 9$$

$$\text{SIM}(Q, Q) = 1 \cdot 1 + 0 \cdot 0 + 0 \cdot 0 + 0 \cdot 0 + 3 \cdot 2 = 10$$

$\vec{a} = (a_1, a_2) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0.0 & 1.0 \\ 0.0 & 0.0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

- CALCOLIAMO LA SIMILARITÀ DEL COSSENO

$$\|\vec{D}_1\| = \sqrt{7^2 + 5^2 + 1^2} = \sqrt{73} \quad \text{GUARDA } \vec{D}_1$$

$$\|\vec{D}_2\| = \sqrt{4^2 + 3^2} = \sqrt{25} = 5 \quad \text{GUARDA } \vec{D}_2$$

$$\|\vec{Q}\| = \sqrt{1^2 + 2^2} = \sqrt{5} \quad \text{GUARDA } \vec{Q}$$

$$\text{COSIM } (\vec{D}_2, \vec{Q}) = \frac{\vec{D}_2 \cdot \vec{Q}}{\|\vec{D}_2\| \|\vec{Q}\|} = \frac{10}{5\sqrt{5}} = 0,894$$

$$\text{COSIM } (\vec{D}_1, \vec{Q}) = \frac{\vec{D}_1 \cdot \vec{Q}}{\|\vec{D}_1\| \|\vec{Q}\|} = \frac{8}{\sqrt{73} \sqrt{5}} \approx 0,665$$

$\vec{D}_1 = \langle \text{AUTOMOBILE IN COLA PER VIA DEL TRAFFICO, BLOCCO TRAFFICO} \rangle$

$\vec{D}_2 = \langle \text{SEMAFORO ROSSO E TRAFFICO INTENSO} \rangle$

$Q = \langle \text{TRAFFICO INTENSO} \rangle$

CALCOLARE LA SIM TRA α e D_1 , $e D_2$

- CALCOLIAMO LE BAG OF WORDS

$D_1 = (\text{AUTOMOBILE}: 1, \text{CODA}: 1, \text{TRAFFICO}: 2, \text{BLOCCHI}: 1)$

$D_2 = (\text{SEMAFORO}: 1, \text{ROSSO}: 1, \text{TRAFFICO}: 1, \text{INTENSO}: 1)$

$Q = (\text{TRAFFICO}: 1, \text{INTENSO}: 1)$

	AUT	CODA	TRAF	BLOC	SEM	ROSSO	INT
D_1	1	1	2	1	0	0	1

D_2	0	0	1	0	1	1	1
-------	---	---	---	---	---	---	---

Q	0	0	1	0	0	0	1
-----	---	---	---	---	---	---	---

$$\|D_1\| = \sqrt{1^2 + 1^2 + 2^2 + 1^2} = \sqrt{7}$$

$$\|D_2\| = \sqrt{1^2 + 1^2 + 1^2 + 1^2} = \sqrt{4} = 2$$

$$\|Q\| = \sqrt{1^2 + 1^2} = \sqrt{2}$$

$$\cos(\alpha, Q) = \frac{1 \cdot 0 + 1 \cdot 0 + 2 \cdot 1 + 1 \cdot 0 + 0 \cdot 0 + 0 \cdot 0 + 0 \cdot 1}{\sqrt{7} \cdot \sqrt{2}} = \frac{2}{\sqrt{14}}$$

$|D_1| \cdot |Q|$

$$\frac{2}{\sqrt{2} \cdot \sqrt{2}} = 0,534$$

$$\cos(\theta_2, \alpha) = \frac{2}{2\sqrt{2}} = 0,707$$

(2) LAVAGNA 26/10/2020

SIANO DATI IN INPUT LA SEGUENTE QUERY Q E IL

DOCUMENTO D:

Q = "INFORMATION RETRIEVAL"

D = "INFORMATION RETRIEVAL AND TEXT RETRIEVAL"

CALCOLARE LA SIMILARITA' DEL COSCENO TRA LA QUERY Q E IL DOCUMENTO

D, ASSUMENDO CHE:

- IL TERMINE "AND" SIA UN STOPWORD.

- IL DOCUMENT FREQUENCY DEI TERMINI "INFORMATION", "RETRIEVAL" E

"TEXT" SIANO RISPETTIVAMENTE 10, 50, 100

- IL NUMERO DEI DOCUMENTI ALL'INTERNO DELLA COLLEZIONE SIA $N = 1000$
- SIA UTILIZZATO IL IDF-IDF COME SCHEMA DI PESATURA DEI TERMINI NEL DOCUMENTO E NELLA QUERY

RAPPRESENTO Q E D SOTTO FORMA DI BAG OF WORDS

$$Q = \langle \text{INFORMATION}: 1, \text{RETRIVAL}: 1 \rangle$$

$$D = \langle \text{INFORMATION}: 1, \text{RETRIVAL}: 2, \text{TEXT}: 1 \rangle$$

	INFORMATION	RETRIVAL	TEXT
Q	1	1	0
D	1	2	1

LOGARITMO IN BASE 10 DEL RAPPORTO TRA IL NUMERO DEI DOCUMENTI E IL DOCUMENT E IL DOCUMENT FREQUENCY

= CALCOLIAMO GLI IDF ↑

$$\text{IDF}_{\text{INFORMATION}} : \log_{10} \frac{1000}{10} = \log_{10} 100 = 2$$

$$\text{IDF}_{\text{RETRIVAL}} : \log_{10} \frac{1000}{20} = \log_{10} 50 = 1,3$$

$$IDF_{TEXT} = \log_{10} \frac{1000}{100} = \log_{10} 10 = 1$$

DEFINIAMO LA MATERIALE TERMINI - DOCUMENTI CHE RAPPRESENTA IL TF-IDF

	\vec{Q}	\vec{D}
INFORMATION	$1 \times 2 = 2$	$1 \times 2 = 2$
RETRIEVAL	$1 \times 1,3 = 1,3$	$2 \times 1,3 = 2,6$
TEXT	0	$1 \times 1 = 1$

$$\vec{Q} = (2, 1, 3, 0) \quad \vec{D} = (2, 2, 6, 1)$$

$$\cos(\vec{D}, \vec{Q}) = \frac{\vec{D} \cdot \vec{Q}}{\|\vec{D}\| \|\vec{Q}\|} = \frac{2 \cdot 2 + 1,3 \cdot 2,6 + 0 \cdot 1}{\sqrt{2^2 + 2,6^2 + 1^2} \sqrt{2^2 + 1,3^2 + 0^2}} = 0,8$$

(?) $1000 \times 100 \times 100 \times 100$

(3) L'HVAGNA 10/11/20

SIA Q UNA QUERY CHE HA 6 DOCUMENTI RILEVANTI NELLA COLLECTIONS. SUPPONIAMO CHE UN

ALGORITMO DI RITROVAMENTO RIPORTI IL SEGUENTE RANKING R_q (R INDICA CHE IL

DOCUMENTO È RILEVANTE, N INDICA CHE IL DOCUMENTO NON È RILEVANTE).

RISULTATO DELLA LISTA È IL TOP DELLA LISTA):

$$R_q: R R R N N N N R N R$$

a) FORNIRE LA DESCRIZIONE SINTETICA DELLE METRICHE: PRECISION, RECALL E

AVERAGGIO PRECISION

b) CALCOLARE PRECISION, RECALL E AVERAGE PRECISION PER LA QUERY Q

c) RIPORTARE LA CURVA DI PRECISION RECALL PER LA QUERY Q, USANDO 11 LIVELLI

STANDARD DI RECALL

$$P \rightarrow \frac{\# \text{DOCUMENTI RILEVANTI RITROVATI}}{\# \text{DOCUMENTI RITROVATI}} = \frac{5}{10} = 0,5 \quad 1$$

$$R = \frac{\# \text{DOCUMENTI RITROVATI RILEVANTI}}{\# \text{DOCUMENTI RILEVANTI}} = \frac{5}{6}$$

DOCUMENTI RILEVANTI

1) LA PRECISIONE E IL RICALL SONO METRICHE CHE CONSENTONO DI VALUTARE LA CORRETTEZZA E LA COMPLETITÀ DI UN IR. LA PRECISIONE MISURA LA CORRETTEZZA E QUANDO CAPIRE LA PROPORTIONE TRA I DOCUMENTI RESTITUITI QUANTI NE SONO RILEVANTI.

IL RICALL ESPRIME LA PROPORTIONE TRA I DOCUMENTI RILEVANTI RITROVATI E I DOCUMENTI RILEVANTI.

$$AP = \frac{1}{m} \sum_{k=1}^m \text{PRECISION}(p=k)$$

$\pi^m = \# \text{ DOCUMENTI RILEVANTI}$

AVERAGE PRECISION CALCOLA LA BONTÀ DEL RANKING DI UN SISTEMA

$$AP = \frac{1+2/2+3/3+4/8+5/10+0}{6} = \frac{4}{6} = \frac{2}{3}$$

Rq

$\checkmark \rightarrow 1$

X → 2/2

X → 3/3

0

0

0

0

X → 4/8

6

X → 5/10 + 0 PERCHE' SONO 6 DOCUMENTI RILEVANTI MA
L'ULTIMO NON L'HA TROVATO

CURVA DI PRECISIONE & RECALL

POS	P	R	P	R
1 1	1/6	= 0,166	1	0
2 1	2/6	= 0,333	1	0,166
3 1	3/6	= 0,5	1	0,3
4			1	0,5
5			1	0,666
6				
7				

PRATICAMENTE SI CREA UNA
COLONNA DI RECALL CHE VA
DA 0 A 1. SI RIPORTANO
I VALORI DELLE PRECISIONI
CALCOLATE PRIMA

$$\begin{array}{ccccccc}
 8 & 4/0 & 4/6 & 0,666 & 0,5 & 0,7 & 0,833 \\
 & & & | & & & \\
 & & & 0,3 & & & \\
 & & & | & & & \\
 & & & 0,1 & & & \\
 & & & | & & & \\
 & & & 0 & & & \\
 & & & | & & & \\
 & 10 & 5/0 & 5/6 & 0,973 & 1 & \\
 & & & | & & & \\
 & & & 0 & & & \\
 & & & | & & & \\
 & & & 0 & & & \\
 & & & | & & & \\
 & & & 1 & & & \\
 \end{array}$$

APPLICO LA PROCEDURA DI INTERPOLAZIONE PER OTTENERE LA PRECISIONE AGGIUNTIVA

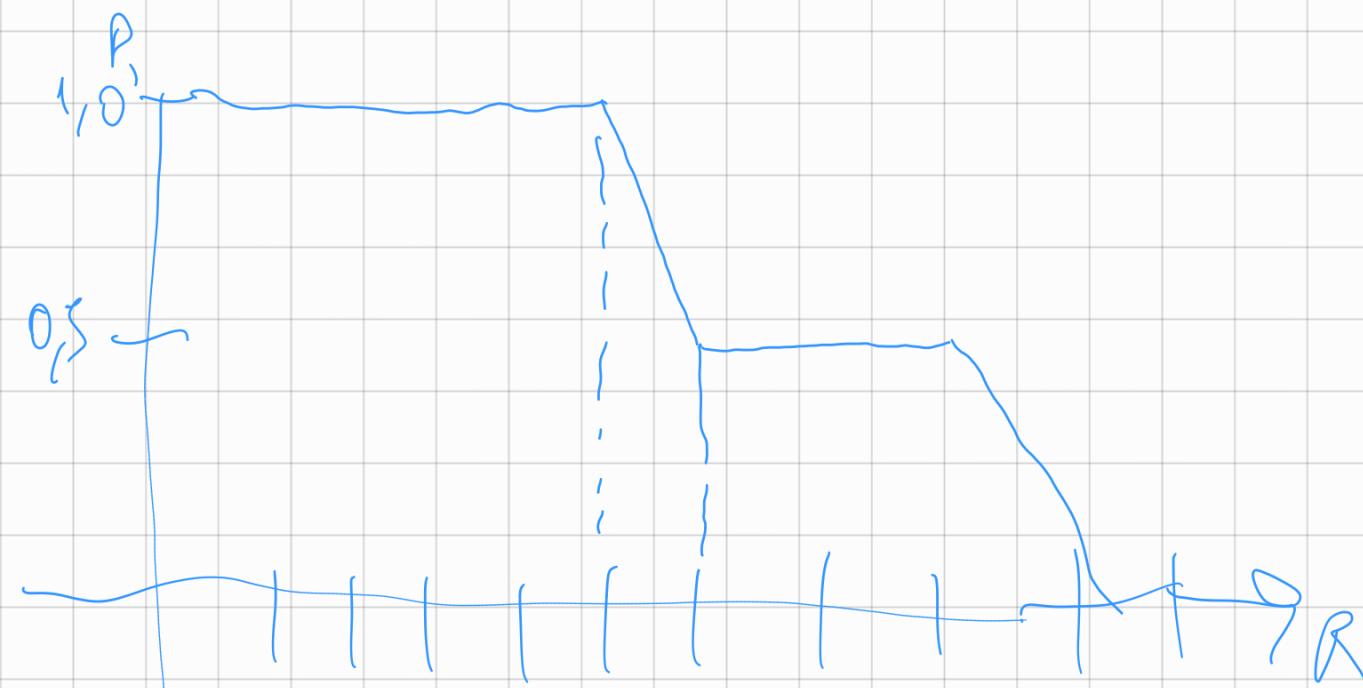
11 PUNTI STANDARD DI RISCAFFOLI

PROCEDURA DI INTERPOLAZIONE

PRECISIONE A LIVELLO r_3 EQUIVALE AL MASSIMO NELL'PRECISIONE PER TUTTI I LIVELLI

$$r \geq r_3$$

$$P(r_3) = \max_{r \geq r_3} P(r)$$



SIA q UNA QUERY CHE HA 6 DOCUMENTI RILEVANTI NELLA COLLEZIONE. SUPONIAMO CHE

UN ALGORITMO DI RI TROVAMENTO APPLICATO A q AI PONTI IL SEGUENTE RANKING R_q :

$D_1, D_5, D_3, D_7, D_9, D_4$

SUPONIAMO CHE D_1, D_3 E D_5 SIANO RILEVANTI PER q .

a) CALCOLARE L'AVG PRECISION PER LA QUERY q , FORMULANDO ANCHE LA DESCRIZIONE DELLA METRICA

b) RIPORTARE LA CURVA DI PRECISION-RECALL PER LA QUERY q , USANDO 11 LIVELLI DI RECALL.

c) SUPONDENDO DI AVERE I GRADII DI RILEVANZA NON LINEARI, E ASSUMENDO CHE D_5 ABbia UN GRADO DI RILEVANZA pari a 3, MENTRE D_1 E D_3 ABBIANO UN GRADO DI RILEVANZA pari a 1. CALCOLA IL VALORE DEL DCG PER q , FORNENDO ANCHE UNA BREVE DESCRIZIONE DELLA METRICA

$$D_1 \quad X \quad AP = 1 + 2/3 + 3/5 =$$

$$D_2 \quad 0 \quad 6$$

$$D_3 \quad X$$

$$D_4 \quad 0$$

$$D_5 \quad X$$

$$D_6 \quad 0$$

INDICE NGL DOCUMENTO RILEVANTE / POSIZIONE NGLLA SEQUENZA

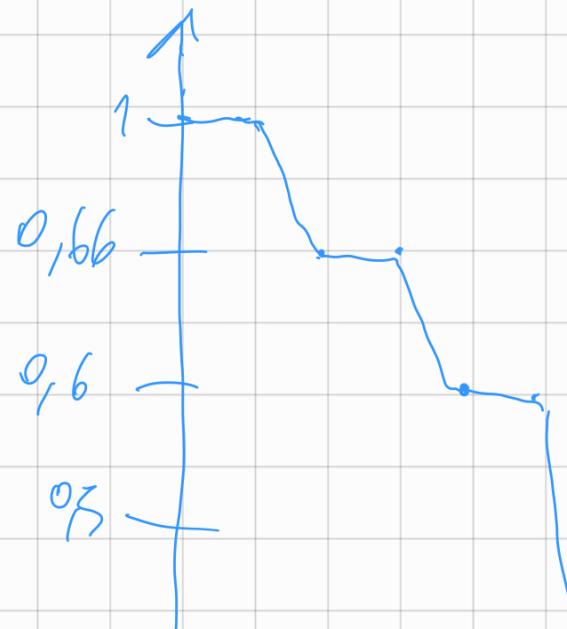
P R → # DOCUMENTO RILEVANTE / # DOCUMENTI RILEVANTI

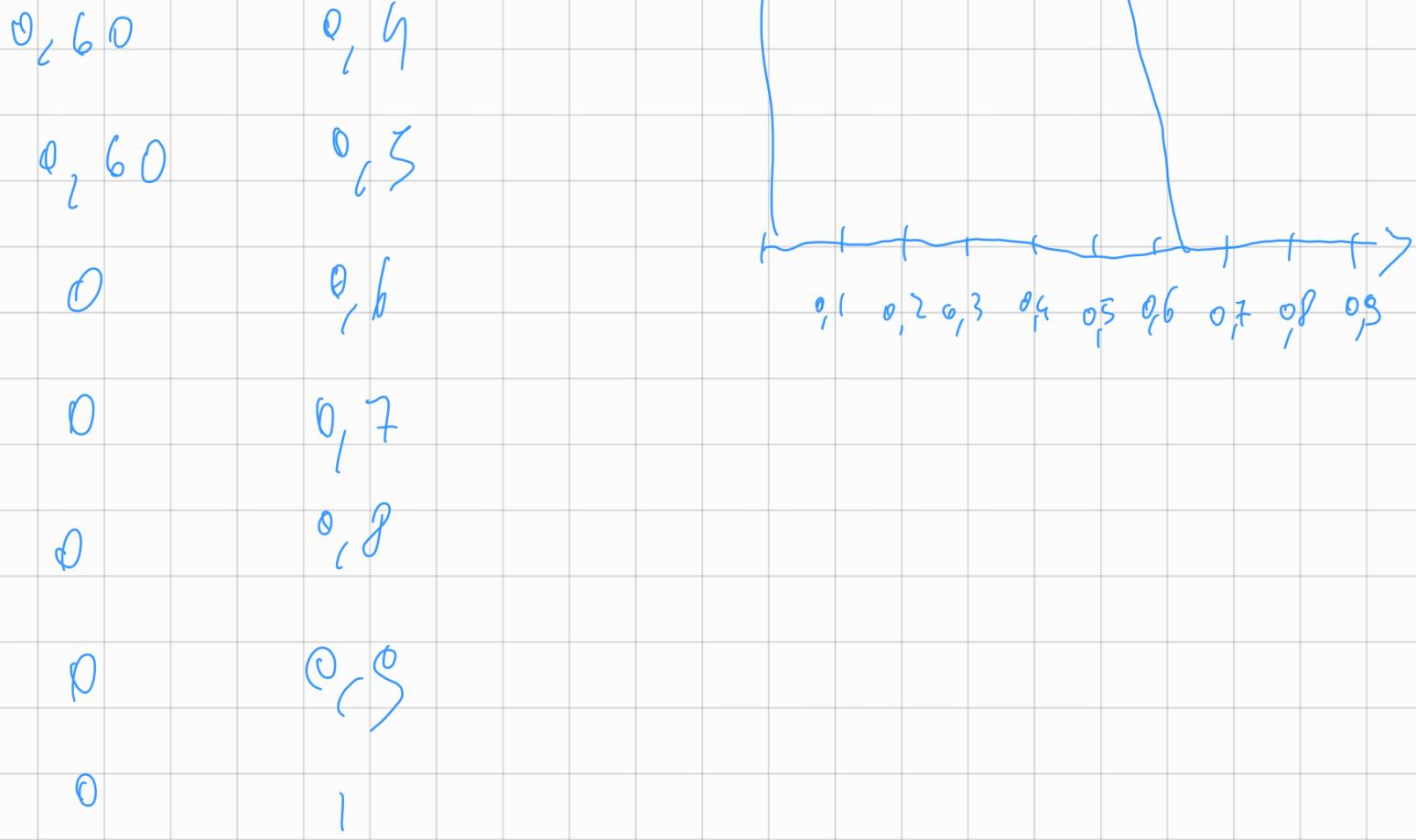
$$1 \quad 1/6$$

$$2/3 \quad 2/6$$

$$3/5 \quad 3/6$$

P	R
1	0
1	0,1
1	0,166
2/3	0,2
0,66	0,3
0,66	0,33





$$DCG[n] = \begin{cases} G[n] & i \geq 1 \\ \frac{G[i]}{\log_2 i} + DCG[i-1] & i \geq 1 \end{cases}$$

$b = (1, 0, 1, 0, 3, 0)$

$$DCG = (1, 0+1, \frac{1}{\log_2 3} + 1, 0+1+\frac{1}{\log_2 3},$$

$$\frac{3}{\log_2 5} + 1 + \frac{1}{\log_2 3}, \frac{3}{\log_2 3} + 1 + \frac{1}{\log_2 3}$$



SI CALCOLA IL VETTORE GAIN PRENDENDO LE RIGUARANZE DEI DOCUMENTI.

SUCCESSIVAMENTE SI COSTRUISCE IL VETTORE DCG CON LA FORMULA

CALCOLIAMO m_{DCG}

SI CONSIDERA IL IDEAL GAIN

$$IG = (3, 1, 1, 900)$$

$$IDCG = (3, \frac{1}{\log_2 2} + 3, \frac{1}{\log_2 3} + 1,$$

$$\frac{1}{\log_2 3} + 4, \frac{1}{\log_2 3} + 4, \frac{1}{\log_2 3} + 4$$

$$n \Delta C G = \frac{\Delta C G}{I \Delta C G}$$



$$R_{q_1} R_{q_2} \cancel{R_{q_3}} = S$$

$$\times \quad 0 \ 1$$

$$0 \quad \times \ 2$$

$$\times \quad \times \ 3$$

$$\times \quad \times \ 4$$

$$0 \quad 0 \ 5$$

$$\times \quad 0 \ 6$$

$$0 \quad \times \ 7$$

$$\times \quad 0 \ 8$$

$$0 \quad 0 \ 9$$

$$0 \quad \times \ 10$$

AVERAGE PRECISION

$$AP_1 = \underbrace{\frac{1}{1} + \frac{2}{2} + \frac{3}{3} + \frac{4}{4} + \frac{5}{5}}_{5}$$

$$AP_2 = \underbrace{\frac{1}{2} + \frac{2}{3} + \frac{3}{4} + \frac{4}{5} + \frac{5}{10}}_{5}$$

Q1

P	R	P	R
1	1/3		0/1
2/3	2/3		0/2
3/4	3/5		0/3
4/6	4/5		0/4
5/8	5/5		0/5

q_2

0,6

0,7

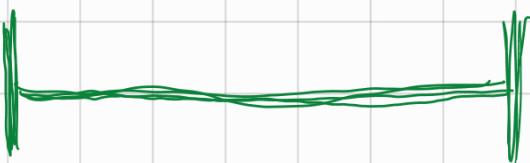
VENERGÈ ESEMPIO

0,8

PRECEDENTI

0,8

{



SI CONSIDERI UNA COLLEZIONE DI 700 DOCUMENTI E DUE QUERY q_1 E q_2 . SI

SUPPONGA CHE:

- q_1 HA COME RISULTATO UN INSIEME DI 70 DOCUMENTI, DI CUI 38 RILEVANTI

- q_2 HA COME RISULTATO UN INSIEME DI 50 DOCUMENTI, DI CUI 25

RILEVANTI

QUALE QUERY CONVIENE FORMULARE? PERCHÉ?

SE SI SAPESSE ANCHE CHE PER OGNI DOCUMENTO DUE QUERY L DOCUMENTI

RILEVANTI NELLA COLLEZIONE SONO 80 LA RISPOSTA POTREBBE CAMBIARE?

PERCHÉ

NON HO NESSUNA INFORMAZIONE SULL'ORDINAMENTO DEI DOCUMENTI DEI RISULTATI
QUINDI NON POSSO VALUTARE IL RANKING, CALCOLO QUINDI LA PRECISIONE E IL
RECALL

$$\text{PRECISIONE} = \frac{\# \text{ DOCUMENTI RILEVANTI RITROVATI}}{\# \text{ DOCUMENTI RITROVATI}}$$

$$P_{q_1} = \frac{38}{70} = 0,54 \quad P_{q_2} = \frac{25}{50} = 0,5$$

CONVIENE q_1 , PERCHÉ HA UNA PRECISIONE PIÙ ALTA

$$\text{RECALL} = \frac{\# \text{ DOCUMENTI RILEVANTI RITROVATI}}{\# \text{ DOCUMENTI RILEVANTI}}$$

$$R_{q_1} = \frac{38}{90} = 0,42 \quad R_{q_2} = \frac{25}{50} = 0,5$$

SARanno quindi i documenti rilevanti per ciascuna query passo calcolare

SIENNO ORA + DOCUMENTI RICEVUTI PER CLASSE D'OGGI, 10-30 - 11-00

ANCHE IL RECALL. BASANDO MI SUL ESSO, CONVIENE SEMPRE LA QUERY

Q1



DESCRIVERE I PROBLEMI DI POLISEMIA E DI SINONIMIA, E SPIEGARE CHE INFLUENZA

HANNO SULLE METRICHE DI PRECISIONE E DI RICHIAMO

LA POLISEMIA, NEL LINGUAGGIO NATURALE, CONSISTE NEZZE PAROLE CHE HANNO PIÙ DI UN

SIGNIFICATO. LA SINONIMIA SONO TERMINI DIVERSI CHE HANNO LO STESSO SIGNIFICATO.

SE IN UNA QUERY Ho UN TERMINE POLISEMICO, SI RIDUCE LA PRECISIONE. LA

SINONIMIA RIDUCE IL RICHIAMO.



SIA DATO IL DOCUMENTO D CON I SEGUENTI TERMINI E RELATIVI OCCORRENZE

a) ESAME^o 4

b) GESTIONE^o 2

c) CONOSCENZA^o 1

SI ASSUMA DI AVERE UNA COLLEZIONE DI 5000 DOCUMENTI IN CUI:

a) IL NUMERO DI DOCUMENTI CONTENENTI IL TERMINE ESAME È 800;

b) IL NUMERO DI DOCUMENTI CONTENENTI IL TERMINE GESTIONE È: 400;

c) IL NUMERO DI DOCUMENTI CONTENENTI IL TERMINE CONOSCENZA È: 10;

CALCOLARE PER OGUNO DEI TRE TERMINI IN d) IL CORRISPONDENTE VALORE

tf - idf.

$$idf_{ESAME} = \log_{10} \left(\frac{5000}{800} \right) = 0,79$$

$$idf_{GESTIONE} = \log_{10} \left(\frac{5000}{400} \right) = 1,09$$

$$idf_{CONOSCENZA} = \log_{10} \left(\frac{5000}{10} \right) = 2,69$$

$$w_{tf, ESAME, d} = \begin{cases} 1 + \log_{10} \frac{tf_{ESAME, d}}{d} & \text{SE } tf > 0 \\ 0 & \text{ALTROVÉ} \end{cases}$$

$$= 1 + \log h = 1,60$$

$$W_h^{\text{GESTIONE}, d} = \begin{cases} 1 + \log t_f^{\text{gestione}, d} & \text{SE } t_f > 0 \\ 0 & \text{ALTRUO} \end{cases}$$

$$= 1 + \log (2) = 1,30$$

$$W_h^{\text{CONOSCENZA}, d} = \begin{cases} 1 + \log t_f^{\text{conoscenza}, d} & \text{SE } t_f > 0 \\ 0 & \text{ALTRUO} \end{cases}$$

$$= 1 + \log (1) = 1$$

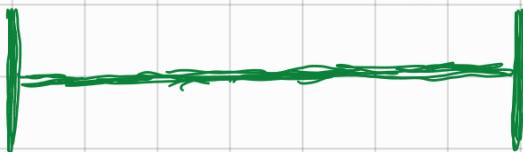
$$T_{f-\text{ioff}}^{\text{ESAME}} = t_f^{\text{(esame, d)}} \cdot \text{ISF}_{\text{(esame, d)}} =$$

$$1,60 \cdot 0,78 \approx 1,264$$

$$T_{f-\text{ioff}}^{\text{GESTIONE}, d} = t_f^{\text{(gestione, d)}} \cdot \text{ISF}_{\text{(gestione, d)}} =$$

$$1,30 \cdot 1,03 \approx 1,417$$

$$T_{f-\text{ioff}}^{\text{CONOSCENZA}} = 2,69 \cdot 1 = 2,69$$



SIANO DATI I SEGUENTI DOCUMENTI ESTRATTI DA UNA COLLEZIONE DI 100 DOCUMENTI:

DOCUMENTI:

$D_1 = "t_1 t_2 t_1 t_3"$

$D_2 = "t_3 t_4"$

$D_3 = "t_1 t_5 t_4 t_6"$

e) FORNIRE LA RAPPRESENTAZIONE DEI DOCUMENTI SOTTO FORMA DI BAG OF WORDS

b) COSTRUIRE L'INDICE INVERTITO DEI DOCUMENTI

c) CALCOLARE LA RAPPRESENTAZIONE TF-IDF PER I PRIMI 3 DOCUMENTI (USARE IL NUMERO DI OCCORRENZE NON NORMALIZZATO PER IL TF)

d) UTILIZZANDO LA SIMILARITÀ DEL COSENZO, DEFINIRE IL RANKING DEI DOCUMENTI

IN RISPOSTA ALLA QUERY q

$q = t_1 t_2 t_1 h$

$Q = \{1, 9, 2, 1, 6\}$

c) BOW $D_1 = \langle T_1 : 2, T_2 : 1, T_3 : 1 \rangle$

$D_2 = \langle T_3 : 1, T_4 : 1 \rangle$

$D_3 = \langle T_1 : 1, T_5 : 1, T_4 : 2 \rangle$

b) INDICE INVERTITO $T_1 \rightarrow D_1, D_3$

$T_2 \rightarrow D_1$

$T_3 \rightarrow D_1, D_2$

$T_4 \rightarrow D_2, D_3$

$T_5 \rightarrow D_3$

c) RAPPRESENTAZIONE TF-IDF

PER CALCOLARE L'IDF HO BISOGNO DI CALCOLARE LA CARDINALITÀ DELLA

COLLEZIONE CHE EQUIVALE A 100 E I DOCUMENT FREQUENCY DEI TERMINI, CIÒ È

QUANTI DOCUMENTI CONTENGONO QUEL TERMINE. NON HO TUTTE LE INFO PER

CALCOLARE L'IDF, QUINDI IPOTIZZO CHE %

$$d(t_{t_1}) = 50 \quad DFT_{T_1} = \log\left(\frac{100}{50}\right) = 0,30$$

$$d(t_{t_2}) = 30 \quad DFT_{T_2} = \log\left(\frac{100}{30}\right) = 0,30$$

$$d(t_{t_3}) = 80 \quad DFT_{T_3} = \log\left(\frac{100}{80}\right) = 0,045$$

$$d(t_{t_4}) = 80 \quad DFT_{T_4} = \log\left(\frac{100}{80}\right) = 0,045$$

$$d(t_{t_5}) = 10 \quad DFT_{T_5} = \log\left(\frac{100}{10}\right) = 1$$

$$d(t_{t_6}) = 30 \quad DFT_{T_6} = \log\left(\frac{100}{30}\right)$$

	T_1	T_2	T_3	T_4	T_5	T_6
D_1	2,030	0,30	0,045	0	0	0
D_2	0	0	0,045	0,045	0	0
D_3	0,30	0	0	2,030	1	0

d) $Q = \langle T_1, T_2, T_6 \rangle$

$$\begin{matrix} T_1 & T_2 & T_3 & T_4 & T_5 & T_6 \\ \log 2 & \log_2 0 & 0 & 0 & 0 & \log \frac{10}{3} \end{matrix}$$

$$\text{COSIM}(D_1, Q) = 2 \log 2 \cdot \log 2 + \log_2 \cdot \log_2$$

$|B_1| |a|$

$$|B_1| = \sqrt{(\log 2)^2 + (\log 2)^2 + \log \frac{10}{3}}$$

$$|q| = \sqrt{(\log 2)^2 + (\log 2)^2 + \log \left(\frac{10}{3}\right)^2}$$



F

DESCRIVERE LA METRICA mDCG (NORMALIZED DISCOUNTED CUMULATIVE GAIN), ILLUSTRANDONE

CALCOLO E PRINCIPI



L'mDCG È UNA METRICA CHE VIENE UTILIZZATA PER VALUTARE L'ACCURATEZZA DEL SISTEMA DI

ALTRIVAMENTO DELL'INFORMAZIONE QUANDO SI HANNO ANCHE NEI GIUDIZI DI GRADIMENTO CHE SONO

IN UNA SCALA DISCRETA. IN QUESTA METRICA I DOCUMENTI RILEVANTI VAZIONO DI PIÙ DEI

DOCUMENTI NON RILEVANTI E SE UN DOCUMENTO RILEVANTE COMPARA IN POSIZIONI PIÙ BASSI

NEL RANKING HA MENO UTILITÀ DI UN DOCUMENTO CHE COMPARA NELLE PRIME POSIZIONI. PER

CALCOLARE QUESTA METRICA, CONSIDERO IL VETTORE DEL GAIN, OVVERO IL VETTORE CHE

ESPRIME L'UTILITÀ DI OGNI SINGOLO DOCUMENTO ALL'INTERNO DEL RESULT SET. PER OTTENERE

IL CUMULATIVO GAIN, LA FORMULA È $\sum_{i=1}^P \frac{\text{REL}_i}{\log(i+1)}$ DOVE REL_i È LA RILEVANZA

$\text{DCG}_2(x)$

DAL DOCUMENTO IN POSIZIONE 2 NELL'VECTORE E' LA POSIZIONE PER CALCOLARE INFINE

L' $m\text{DCG}$, SI CALCOLA L'IDEAL DCG, IL CUI CALCOLO E' SIMILE AL PROCEDIMENTO SOPRA ELENCATO

SOLO CHE SI CALCOLA SULLA BASE DEL GAIN ORDINATO IN MANIERA OTTIMALE E SUCCESSIVAMENTE SI

CALCOLA $\frac{\text{DCG}}{m\text{DCG}}$



SIA Q UNA QUERY CHE HA 5 DOCUMENTI RILEVANTI NELLA COLLEZIONE. SUPPONIAMO CHE UN

ALGORITMO DI RITROVAMENTO APPLICATO A Q RIPORTI IL SEGUENTE RANKING $R_q : D_1, D_5,$

D_3, D_7, D_9, D_4 . SUPPONENDO DI AVERE GLI GRUNZI DI RILEVANZA BINARI ESPRESI

IN UNA SCALA A 5 VALORI (1-5), E ASSUMENDO CHE D_1 E D_9 ABBIANO RILEVANZA

PARI A 5, MENTRE D_5 ABbia RILEVANZA PARI A 3, CALCOLARE IL VALORE DELL' $m\text{DCG}$

PER q.

$$R_q = D1 \ D5 \ D3 \ D7 \ D9 \ D4$$

$$b = (5, 3, 0, 0, 5, 0)$$

$$DCG[i] = \begin{cases} b[i] & i=1 \\ \frac{b[i]}{\log_2 i} + DCG[i-1] & i>1 \end{cases}$$

$$DCG = \left(5, \frac{3}{\log_2 2} + 5, \frac{8}{\log_2 3}, \frac{8}{\log_2 4}, \frac{5}{\log_2 5} + 8, \frac{5}{\log_2 6} + 8 \right)$$

↓
8

$$DCG = (3, 8, 8, 8, 10.2, 10.2)$$

$$IG = (55, 3, 0, 0)$$

$$IDCG = \left(5, \frac{5}{\log_2 2} + 5, \frac{3}{\log_2 3} + 10, \frac{3}{\log_2 4} + 10, \frac{3}{\log_2 5} + 10, \frac{3}{\log_2 6} + 10 \right)$$

$$\approx (5, 10, 11.8, 11.8, 11.8, 11.8)$$

$$n \overline{DCB} \rightarrow \frac{\overline{DCB}}{\overline{IDCB}} = \left(\frac{5}{5}, \frac{8}{60}, \frac{8}{11.8}, \frac{8}{11.8}, \frac{8}{11.8} \right)$$



DESCRIVERE IN MANIERA SINTETICA I PRINCIPI ALLA BASE DEL PAGE RANK, FOCALIZZANDO L'ATTENZIONE SULLA FORMULAZIONE BASATA SUL FLOW MODEL E SULLE MATRICI DI ADIACENZA STOCASTICA

$$\begin{matrix} V_1 \\ \vdots \end{matrix}$$

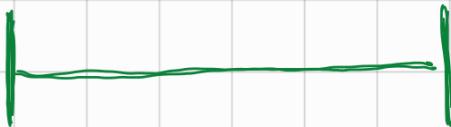
IL PAGE RANK È UN ALGORITMO CHE PERMETTE DI ASSEGNARE UNO SCORE DI IMPORTANZA AGLI PAGINE IN BASE ALLA TIPOLOGIA DEL GRAFO CHE FORMANO LE CONNESSIONI E OGNI PAGINA PUÒ ESSERE COLLEGATA AD ALTRE PAGINE. OGNI LINK ENTRANTE È VISTO COME UN VOTO, SE UNA PAGINA HA PIÙ LINK ENTRANTI, OTTIENE IMPORTANZA DA DIVERSE PAGINE, AUMENTANDO COSÌ LA SUA IMPORTANZA. SULLA BASE DEL FLOW MODEL, L'IMPORTANZA DI UNA PAGINA j È DATA DALLA SOMMA DEL RANK DI TUTTE LE PAGINE i CHE PUNTANO A j FRATTO L'OUTDEGREE, CIOÈ SE UNA PAGINA HA UN LINK VERSO TANTE ALTRE PAGINE, DISTRIBUISCE UN'IMPORTANZA INVERSA E PROPORZIONALE AL NUMERO

DI LINK. FACENDO UNA SERIE DI EQUAZIONI, CHE MODELLANO IL FLOSSO DI LINK, E
 AGGIUNGENDO UN'EQUAZIONE CHE STABILISCE CHE LA SOMMA DI TUTTI I RANK DEVE
 ESSERE UGUALE A 1, HO UN SISTEMA DI EQUAZIONI, LA CUI SOLUZIONE GENERA
 IL PAGE RANK DELLE PAGINE. IL PROBLEMA È CHE QUANDO HO MOLTE PAGINE,
 IL CALCOLO RISULTA INFATTIBILE PERCHÉ AURO' UN NUMERO ELEVATO DI
 PAGINE. ALLORA POTREMO MODELLARE IL GRAFO IN UNA MATRICE DI ADIACENZA
 STOCASTICA E QUINDI SE UNA PAGINA i PUNTA A UNA PAGINA j ALLORA m_{ij} SARÀ
 $1/d_i$,cioè $1/(\text{outdegree della pagina } i)$, ALTRIMENTI È PARI A 0. A QUESTO PUNTO
 LE EQUAZIONI POSSONO ESSERE RISCRITTE NELLA FORMA $r = m \cdot r$. QUESTA FORMA SI
 RISOLVE ATTRAVERSO IL METODO DELLE POTENZE. PARTENDO DA UNA CONFIGURAZIONE
 INIZIALE DEL VETTORE DI RANK, DISTRIBUITO A PIACIMENTO, E' LTERO MOLTIPLICANDO
 $m \cdot r$ TANTE VOLTE FINO A QUANDO IL VETTORE NON SI "NORMALIZZA", cioè FINO A QUANDO
 LA DIFFERENZA TRA UN'ITERAZIONE E LA SUCCESSIVA NON È PICCOLA. ALLA FINE OTTENGO
 UN VALORE DEL VETTORE r DI RANK CHE NON CAMBIA

$$r_j = \sum_{i \in S} \frac{r_i}{d_i} \quad r_j \Rightarrow \text{RANK DELLA PAGINA } j$$

$$r_i \Rightarrow \text{RANK DELLA PAGINA } i \text{ CHE}$$

$\delta(i \Rightarrow)$ LINI USCENTI DALLA PAGINA I



DESCRIVERE IL PROCESSO DI MODIFICA DELLA QUERY BASATO SUL METODO DEL RELEVANCE FEEDBACK (ALGORITMO DI ROCCHIO) V. 1

FEEDBACK (ALGORITMO DI ROCCHIO) V. 1

ALLA BASE DI QUESTO METODO VI È UNA COLLEZIONE DI DOCUMENTI RILEVANTI E NON

RILEVANTI, SE LO CONOSSO PER UNA QUERY QUALI SONO I DOCUMENTI OTTIMALI E NON

OTTIMALI, POTREI FORMULARE UNA QUERY OTTIMALE CHE DIVIDE LA COLLEZIONE IDENTIFICANDO

I DOCUMENTI RILEVANTI DA QUELLI NON RILEVANTI. LA QUERY OTTIMALE È DATA DAL

CENTROIDE DEI DOCUMENTI RILEVANTI MENO IL CENTROIDE DEI DOCUMENTI NON RILEVANTI.

IL PROBLEMA È CHE QUANDO HO UNA QUERY NON CONOSCO LA COLLEZIONE, QUINDI

CREO UNA QUERY INIZIALE q_0 , E DOPO AVER OTTENUTO DEL FEEDBACK DA PARTE

DELL'UTENTE, PER POI RAPPINERÈ QUESTA QUERY TRAMITE UNA COMBINAZIONE LINEARE

TRA LA QUERY DI PARTENZA E LA QUERY MODIFICATA.

$\alpha + \beta$

QUESTA QUERY MODIFICATA NON FA ALTRO

90 CENTRALIZZARE DOCUMENTI RILEVANTI
DOCUMENTI NON RILEVANTI CHE MODIFICARE IL VETTORE QUERY DEI

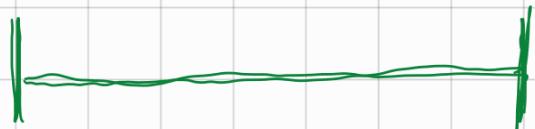
DOCUMENTI RILEVANTI E ALLONTANARLO DAI DOCUMENTI NON RILEVANTI. I PESI

α, β, γ DANNNO UN PESO ALLA QUERY INIZIALE AI DOCUMENTI RILEVANTI E AI DOCUMENTI

NON RILEVANTI. IN GENERE I DOCUMENTI RILEVANTI SI PESANO DI PIÙ RISPETTO AI

DOCUMENTI NON RILEVANTI. I PRIMI HANNO UN PESO DI 0,75 E I SECONDI UN PESO DI 0,25.

SE FACENDO LA DIFFERENZA I PESI DIVENTANO NEUTRI, VENGONO POSTI A 0



SIA q UNA QUERY CHE HA 6 DOCUMENTI NELLA COLLEZIONE. SUPPONIAMO CHE UN ALGORITMO DI

RITROVAMENTO APPLICATO A q RIPORTI IL SEGUENTE RANKING

$R_q: D_1 \ D_2 \ D_3 \ D_4 \ D_5 \ D_6$

SUPPONIAMO CHE $D_2, D_4 \in D_6$ DOCUMENTI RILEVANTI PER q

2) CALCOLARE L' AVERAGE PRECISION E IL RECALL PER LA QUERY q

FORNENDO UNA DESCRIZIONE DELLE METRICHE

b) RIPORTARE LA CURVA DI PRECISIONE-RECALL PER LA QUERY q_1 , USANDO GLI 11

LIVELLI STANDARD DI RECALL

$R = D_1 \ D_2 \ D_3 \ D_4 \ D_5 \ D_6$

0 X 0 X 0 X

$$AVP = \frac{1/2 + 2/4 + 3/6}{6} = 0,25$$

$$RECALL = \frac{\# DOC RILEVANTI}{\# DOC RITROVATI} = \frac{3}{6}$$

0

$$X \quad P = 0,5 \quad R = 1/6 = 0,167$$

0

$$R = X \quad P = 0,5 \quad R = 2/6 = 0,33$$

P

$$X \quad P = 0,5 \quad R = 3/6 = 0,5$$

P

0,5

R

0

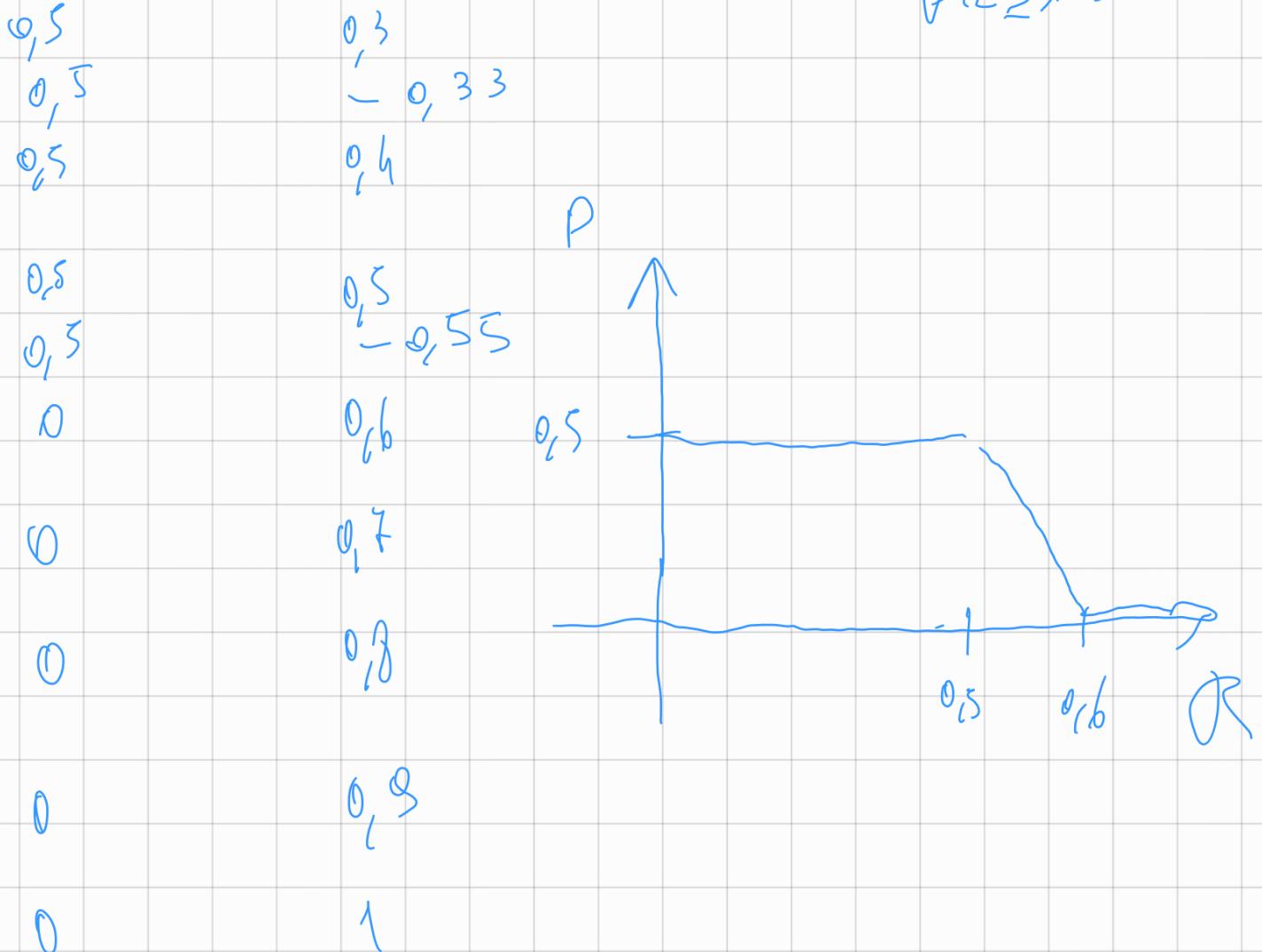
0,5

0,5

0,1

~0,167

$$P() = R()$$



ESERCIZI COLLABORATIVE FILTERING

SIA DATA LA SEGUENTE MATRICE UTENTI-ITEM DI UN SISTEMA DI FILTRAGGIO COLLABORATIVO, I CUI RATING DI GRAVIMENTO SONO ESPRESSSI IN UNA SCALA DISCRETA DA 1 A 5.

	I ₁	I ₂	I ₃	I ₄	I ₅
U ₁	2	4	1	?	1
V ₂		5	2		
U ₂	3	5	1	?	

CORRISPONDENTI ALLA
LAUAGNA 6

U ₁	1	4	4	2	2
----------------	---	---	---	---	---

CALCOLARE LA PREDIZIONE PER L'ITEM I₄ PER L'UTENTE ATTIVO U₁, UTILIZZANDO UN ALGORITMO DI USER-TO-USER COLLABORATIVE FILTERING, UNA NEIGHBORHOOD SIZE PARI A 2 E IL COSENCO COME MISURA DI SIMILARITÀ TRA GLI UTENTI.

PER RISOLVERE QUESTO ESERCIZIO, DOVREI CALCOLARE LA SIMILARITÀ TRA L'UTENTE U₁ E GLI UTENTI U₂, U₃, U₄, POICHÉ U₂ NON HA ASSEGNAZIO UN VOTO A I₄ NON POTRÀ ESSERE UN NEIGHBOR. QUINDI VADO A CALCOLARE LA SIMILARITÀ TRA U₁ E U₃, E TRA U₁ E U₄:

ALGORITMO USER - USER
COLLABORATIVE FILTERING

$$\text{COSIM}(U_1, U_3) = \frac{2 \cdot 3 + 4 \cdot 5 + 1 \cdot 2}{\sqrt{2^2 + 4^2 + 1^2 + 1^2} \cdot \sqrt{3^2 + 5^2 + 1^2 + 2^2}} = 0,96$$

$$\text{COSIM}(U_1, U_4) = \frac{2 \cdot 4 + 1 \cdot 4 + 1 \cdot 2}{\sqrt{2^2 + 4^2 + 1^2 + 1^2} \cdot \sqrt{4^2 + 4^2 + 2^2 + 2^2}} = 0,47$$

MEDIA DEI VOTI DI U₁

$$\bar{v}_{U_1} = \frac{2+4+1+1}{4} = 2 \quad \bar{v}_{U_3} = \frac{11}{4} = 2,75 \quad \bar{v}_{U_4} = 3$$

PREDIZIONE

$$P(U_1, I_4) = 2 + 0,96(1-2,75) + 0,47(2-3) = 0,50$$

$$0,96 + 0,47$$

UTILIZZANDO L'ALGORITMO ITEM-ITEM COLLABORATIVE FILTERING, DEVO

CALCOLARE LE SIMILARITÀ GLI ITEM.

$$|I_1| = \sqrt{2^2 + 3^2 + 4^2}$$

$$|I_2| = \sqrt{4^2 + 5^2 + 5^2}$$

$$|I_3| = \sqrt{21}$$

$$|I_4| = \sqrt{5}$$

$$|I_5| = \sqrt{9}$$

$$\text{SIM}(I_1, I_1) = \frac{3 \cdot 1 + 4 \cdot 2}{|I_1| \cdot |I_1|} = 0,91$$

$$\text{SIM}(I_1, I_2) = \frac{5 \cdot 1}{|I_1| \cdot |I_2|} = 0,27$$

$$\text{SIM}(I_1, I_3) = \frac{4 \cdot 2}{|I_1| \cdot |I_3|} = 0,78$$

$$\text{SIM}(I_1, I_5) = \frac{1 \cdot 2 + 2 \cdot 2}{|I_1| \cdot |I_5|} = 0,89$$

I DUE NEIGHBOR PIÙ VICINI SONO I_1 E I_5 , CON SIMILARITÀ PIÙ ALTA.

$$P(U_1, I_4) = \frac{2 \cdot 0,91 + 1 \cdot 0,89}{0,91 + 0,89} = 1,51.$$



1) ILLUSTRARE IN MANIERA SINTEtica I SEGUENTI PROBLEMI:

a) OVERSPECIALIZATION NEI RECOMMENDER SYSTEMS DI TIPO
CONTENT-BASED

FATTA

b) GREYSHEEP NEI RECOMMENDER SYSTEM DI TIPO COLLABORATION

2) DESCRIVERE IN MANIERA SINTEtica I CONCETTI FONDAMENTALI A BASE DEI MODELLI RDP, IN PARTICOLARE I CONCETTI DI RISORSA, PROPRIETÀ ESTATEMENT

- 3) DESCRIVERE IL PROBLEMA DELLO SPIDER TRAP E SGL DEAD END NELL'ALGORITMO PAGERANK E ILLUSTRARE UNA POSSIBILE SOLUZIONE
- FATTA \rightarrow
- 4) DESCRIVERE, COMMENTANDO OPPORTUNAMENTE, LA FUNZIONE PER IL CALCOLO DELLE PREZISSIONI DEI RATING IN UN ALGORITMO DI FILTRAGGIO COLLABORATIVO DI TIPO USER TO USER.

SI TRAZZASCI QUESTA PARTE PER PROCEDERE PRIMA CON GLI ESERCIZI



SIA q UNA QUERY I CUI DOCUMENTI RILEVANTI NELLA COLLECTION SONO 5.
 SIANO S_1 E S_2 DUE SISTEMI CHE RIPORTANO I SEGUENTI PRIMI 10 RISULTATI IN RISPOSTA ALLA QUERY q . R INDICA CHE UN DOCUMENTO È RILEVANTE, N INDICA CHE IL DOCUMENTO NON È RILEVANTE

$S_1: R \ N \ R \ N \ N \ R \ N \ N \ R \ R$

$S_2: N \ R \ N \ N \ R \ R \ R \ N \ N \ N$

CALCOLARE L'ACCURATEZZA DEI DUE SISTEMI PER LA QUERY q ,

UTILIZZANDO LE SEGUENTI METRICHE:

$P@1$, $P@5$, $P@10$, R-PRECISION, AVERAGE PRECISION

DEFINIZIONE METRICHE

LA PRECISION@K È LA PRECISIONE CALCOLATA TENENDO CONTO DEI PRIMI K RISULTATI. $P@k = \frac{\# \text{doc RILEVANTI}}{k}$

$$P@1(S1) = 1 \quad P@5(S1) = \frac{2}{5} \quad P@10(S1) = \frac{5}{10}$$

LA R-PRECISION È SEMPRE UNA METRICA CHE CALCOLA LA PRECISIONE A UN CERTO LIVELLO DELLA LISTA, DOVÉ PERÒ LA LISTA VIENE TAGLIATA AL VALORE DI RECALL, CIOÈ AL NUMERO DI DOCUMENTI RILEVANTI.

IN QUESTO CASO, VISTO CHE VI SONO SOLO 5 DOCUMENTI RILEVANTI, LA R-PRECISION CORRISPONDE ALLA PRECISION@5.

$$P@CS2) = 0 \quad P@5(S2) = \frac{2}{3} \quad P@10(S2) = 2/5$$

L'AVGPRECISION È UNA METRICA CHE CONSENTE DI CALCOLARE L'ACCURATEZZA DEL SISTEMA, DOVÈ OLTRE A TENERE IN CONSIDERAZIONE LA PRECISIONE, TIENE IN CONSIDERAZIONE ANCHE LA BONTÀ DEL RANKING.

LA FORMULA DELL'AVGPRECISION È :

$$\frac{\sum_{k=1}^K \text{PRECISION}}{K}$$

$$AVP(S1) = \underbrace{1 + 2/3 + 3/6 + 4/9 + 5/10}_5$$

$$AVP(S2) = \frac{1/2 + 2/5 + 3/6 + 4/7 + 0}{5}$$

CURVA DI PRECISIONE E RECALL

$$R = 4/5 = 0,8$$

$$P = 9/13 = 0,69$$

$S_1: X \quad 0 \quad X \quad 0 \quad 0 \quad X \quad 0 \quad 0 \quad X \quad X$

$$P_{121}$$

$$R = 1/5 = 0,2$$

$$P = 3/6 = 0,5$$

$$R = 3/5 = 0,6$$

$$P = 5/10 = 0,5$$

$$R = 5/5 = 1$$

P R

1 0

1 0,1

1 0,2

0,66 0,3

0,66 0,4

0,5 0,5

0,5 0,6

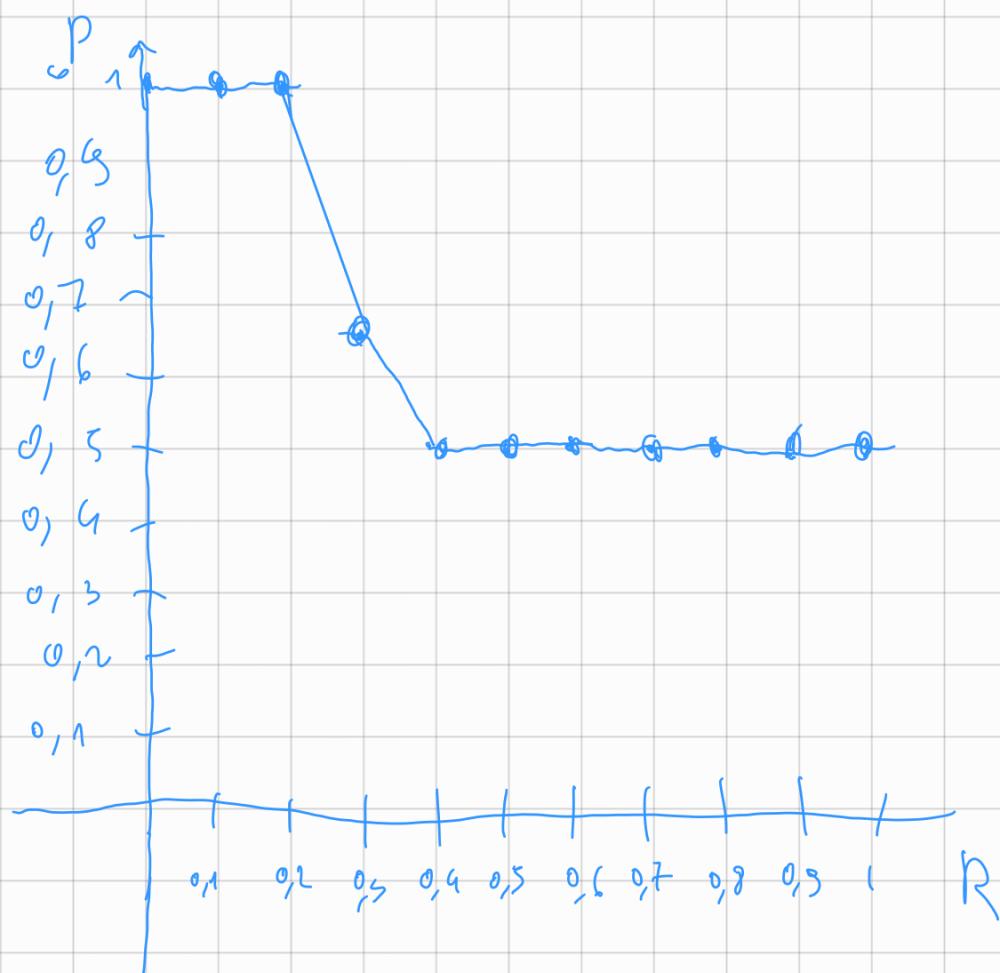
0,5 0,7

$$P(n_3) = \max_{n > n_3} P(n)$$

~~0,5 0,4~~ 0,8 → VA RICALCOLATA

0,5 0,9

0,5 1



DESCRIVERE IL PROCESSO DI MODIFICA DELLA QUERY BASATO
SUL METODO DEL RELEVANCE FEEDBACK V.2



IL PROCESSO DI RELEVANCE FEEDBACK SERVE PER RAFFINARE LE QUERI,
CIOÈ L'OBBIETTIVO È QUELLO DI MIGLIORARE LA QUERY INTEGRANDO IL
FEEDBACK DELL'UTENTE AFFINCHÉ IL SISTEMA POSSA MIGLIORARE LE
SUO PERFORMANZE. SE NOI ABBIAMO UNA COLLEZIONE E PER UNA QUERY
SPECIFICA q

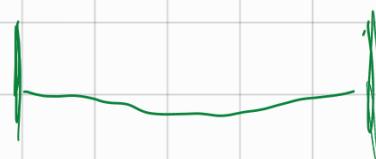
SAPESSIMO ESATTAMENTE QUALI SONO I DOCUMENTI RILEVANTI E I DOCUMENTI NON RILEVANTI, LA QUERY OTTIMEZZE CHE MI PERMETTE DI RITROVARE I DOCUMENTI RILEVANTI ED ESCLUDERE I DOCUMENTI NON RILEVANTI. SAREBBE IL CENTRO IDE DEI DOCUMENTI RILEVANTI MA NO IL CENTRO IDE DEI DOCUMENTI NON RILEVANTI. QUESTA QUERY È SOLO UNA QUERY TEORICA, NON È POSSIBILE FORMULARLA PERCHÉ NON SO QUALI SONO I DOCUMENTI RILEVANTI E NON RILEVANTI, QUINDI PROcedo ad APPROXIMARSI. ESEGUE LA QUERY q_0 SULLA COLLECTIONS, È DILEGO L'UTENTE A FORNIRGÈ DEL FEEDBACK SUI DOCUMENTI RITROVATI, DIVIDENDO I DOCUMENTI RILEVANTI DA QUELLI NON RILEVANTI. LA QUERY MODIFICATA NON È ALTRO CHÉ UNA COMBINAZIONE LINEARE CON DETERMINATI PESI DELLA QUERY INIZIALE.

$$q_0 + \beta_{\text{CENTROIDE}} - \gamma_{\text{CENTROIDE}}$$

\rightarrow

$\beta_{\text{CENTROIDE}}$	$\gamma_{\text{CENTROIDE}}$
DEI DOCUMENTI	DEI DOCUMENTI
RILEVANTI	NON RILEVANTI

QUESTA NUOVA QUERY È UN MIGLIORAMENTO RISPETTO ALLA QUERY INIZIALE. IL VETTORE QUERY INIZIALE SI AUDICINA SEMPRE DI PIÙ AL CENTRO DI MASSA DEI DOCUMENTI RILEVANTI E SI È ACCONTANATO DA QUELLO DEI DOCUMENTI NON RILEVANTI.



ILLUSTRARE IN MANIERA SINTETICA I SEGUENTI PROBLEMI DEI RECOMMENDER SYSTEMS DI TIPO COLLABORATIVO:

- a) COLD-START
- b) GREY SITEP

IL COLD-START, OVVERO PARTENZA A FREDDO, SI VERIFICA QUANDO NELLA MATRICE CI SONO POCHISSIMI RATING ESPRESI DAGLI UTENTI PER GLI OGGETTI, E QUINDI IN QUESTO CASO IL SISTEMA NON RIESCE A CALCOLARE

LE SIMILARITÀ E QUINDI NON RIESCE A FORNIRE LE RACCOMANDAZIONI.

IL GREY-SHEEP È IL PROBLEMA IN CUI ALCUNI UTENTI, INVECE DI POLARIZZARE I PROPRI VOTI, DANNO SEMPRE VOTI INTERMEDII ALL'INTERNO DELLA SCALA, VOTI CHE NON FANNO CAPIRE EFFETTIVAMENTE LE PREFERENZE DEGLI UTENTI, CREANDO PERICOLTÀ NEL TROVARE UTENTI SIMILI.



DESCRIVERE LE TECNICHE D'ERRORE MAE E RMSE UTILIZZATE PER LA VALUTAZIONE DELL'ACCURATEZZA DEI RECOMMENDER SYSTEMS.



LA VALUTAZIONE DEI RECOMMENDER SYSTEM, ATTRAVERSO L'USO DI QUESTE METRICHE D'ERRORE AUGMENTA CON I TASK DI RATING PREDICTION, CIÒ È QUANDO DEVO PREDIRE I RATING ACCINTERNO DI UN ALGORITMO DI COLLABORATIVE FILTERING.

MAE (MIN ABSOLUTE ERROR) È LO SCOSTAMENTO MEDIO CHE SI HA TRA IL RATING EFFETTIVO PRESENTE NEL TEST SET E IL RATING PREDETTO DALL'ALGORITMO DALL'ALGORITMO DI RACCOMANDAZIONE.

$$MAE = \frac{\sum_{u_i \in T} |r_{u_i} - \hat{r}_{u_i}|}{|T|}$$

T = TRAINING SET

r_{u_i} = VOTO DELL'UTENTE U
PER L'ITEM i

\hat{r}_{u_i} = VOTO PREDETTO
DALL'ALGORITMO

RMSÈ (ROOT MIN SQUARE ERROR) HA LO STESSO SIGNIFICATO DEL MAE,

CON L'UNICA DIFFERENZA CHE IN QUESTA METRICA SI EFFETTUÀ LA RADICE QUADRATA DEL QUADRATO DEGLI SCARTI IN MODO DA POTER EVIDENZIARE ERRORI MAGGIORI

$$RMSE = \sqrt{\frac{\sum_{u_i \in GT} (r_{u_i} - \hat{r}_{u_i})^2}{|GT|}}$$



DESCRIVERE, COMMENTANDO OPPORTUNAMENTE, LA FUNZIONE PER IL CALCOLO DELLE PREDIZIONI DEI RATING IN UN ALGORITMO DI FILTRAGGIO COLLABORATIVO DI TIPO USER TO USER



* PRIMA INSERIRE LA PARTE DESCRIPTIVA

$$PRED(a, p) = \bar{r}_a + \frac{\sum_{b \in N} SIM(a, b) \cdot (\hat{r}_{b,p} - \bar{r}_b)}{\sum_{b \in N} SIM(a, b)}$$

\bar{r}_a = MEDIA DEI VOTI DELL'UTENTE a

$SIM(a, b)$ = SIMILARITÀ TRA L'UTENTE a E b

$\hat{r}_{b,p}$ = RATING DELL'UTENTE b ALL'ITEM p

\bar{r}_b = MEDIA DEI VOTI DELL'UTENTE b

DATA UNA MATRICE USER-ITEM, PER CALCOLARE LA PREDIZIONE CHE UN UTENTE HA ASSEGNATO A UN DETERMINATO ITEM SI POTREBBERI OTTENERE UN

MEDIANO TUTTI I GIUDIZI CHE GLI ALTRI UTENTI HANNO ESPRESSO PER QUELL'ITEM
ATTRAVERSO UNA MEDIA ARITMETICA. COSÌ FACENDO PERÒ, CONSIDERAREI ANCHÉ
I GIUDIZI DI UTENTI CHE NON SONO SIMILI IN TERMINI DI PREFERENZE
RISPETTO ALL'UTENTE PER CUI STO CALCOLANDO LA PREDIZIONE. QUINDI PER
MIGLIORARE LA FORMULA, POTREI CALCOLARE UN INSERIMENTO DI NEIGHBOURS E'
CONSIDERARE SOLO LE LORO OPINIONI. POTREI MIGLIORARE ULTERIORMENTE IL
CALCOLO, EFFETTUANDO UNA MEDIA PESATA, NELLA QUALE GLI UTENTI PIÙ SIMILI
VALGONO DI PIÙ RISPETTO A UTENTI MEZZO SIMILI CHE VALGONO DI MENO.

INOLTRE VISTO CHE DIVERSI UTENTI HANNO STILI DI RATING DIVERSI, INVECE
DI FARE LA MEDIA PESATA DEI RATING, FACCIO LA MEDIA PESATA IN ELENCO SCARTO
RISPETTO ALLA MEDIA DI QUESTI RATING, OTTENENDO COSÌ ANCORA UNO SCARTO
CHE VADO A SOMMARE AL VOTO MEDIO DELL'UTENTE ATTIVO



DESCRIVERE IN MANIERA SINTETICA I PRINCIPI ALLA BASE DEL PAGE RANK, FOCALIZZANDO
L'ATTENZIONE SULLA FORMULAZIONE BASATA SUL FLOW MODEL. ✓2

IL PAGE RANK È UN ALGORITMO PER ASSEGNARE UNO SCORE A UNA PAGINA ALL'INTERNO
DI UNA RAPPRESENTAZIONE A GRAFO. SUPponendo che il WEB sia costituito da
UN INSERIMENTO DI NODI CHE SONO LEGATI TRA DI LORO ATTRAVERSO DEGLI ARCHI
QUANDO C'È UN COLLEGAMENTO IPERTEXTUALE, QUELLO CHE IL PAGE RANK FA
È QUELLO DI ANDARE A ANALIZZARE LA TOPOLOGIA DI QUESTO GRAFO. E ANDARE
AD ASSEGNARE AGLI SCORE A DELLE PAGINE. I PRINCIPI DI BASE SONO
CHE SE UNA PAGINA RICEVE IL LINK DA TANTE PAGINE, SIGNIFICA CHE È
UNA PAGINA MOLTO IMPORTANTE. SE RICEVE DEI LINK DA DIVERSE PAGINE
TRASFERISCE UNA CERTA IMPORTANZA. IL MECCANISMO DEL LINK ENTRANTE PUÒ
ESSERE CONSIDERATO COME UN MECCANISMO DI VOTO, TANTI PIÙ LINK INTRANTI SONO
PRESENTI, TANTO PIÙ È IMPORTANTE LA PAGINA. QUANDO UNA PAGINA NON HA
UN'ALTRA NE TRASFERISCE UN PARTITO DELLA SUA IMPORTANZA CHE È INVERSORAMENTE
PROPORTIONALE AL NUMERO DI LINK USCENTI. NEL FLOW MODEL IL RANK DI UNA PAGINA

SI È DATO DA:

$$r_j = \sum_{i \rightarrow j} \frac{r_i}{d_i}$$

$r_i \rightarrow$ PAGINE CHE PUNTANO A J

$d_i \rightarrow$ OUTDEGREE DELLA PAGINA i

$r_j \rightarrow$ RANK DELLA PAGINA J DA CALCOLARE

SE ABBIAMO m PAGINE ALL'INTERNO DEL GRAFO AVREMO m EQUAZIONI CHE MODELLANO IL FLUSSO. SICCOME ABBIAMO m EQUAZIONI IN m INCognITE, PER AVERE SICUREZZA CHE C'È UN'UNICA SOLUZIONE IMPONIAMO CHE LA SOMMA DI TUTTI I RANK SIA PARI A 1; QUINDI AVEREMO UN SISTEMA DEL TIPO:

$$\left\{ \begin{array}{l} r_j = \sum_{i \rightarrow j} \frac{r_i}{d_i} \\ \sum_i r_i = 1 \end{array} \right.$$

AVEREMO QUINDI UN SISTEMA DI m+1 EQUAZIONI A m INCognITE E POSSIAMO RISOLVERLO AD ESEMPIO CON IL MECCANISMO DELLA SOSTITUZIONE, E OTTERNI AMPO IL RANK ASSIGNATO A DETERMINATE PAGINE.

QUESTA FORMULAZIONE HA DEI LIMITI QUANDO SI HANNO MOLTE PAGINE, QUINDI POSSO DARE UNA FORMULAZIONE SOTTO FORMA DI ALGEBRA LINEARE. IL GRAFO VENGÀ MODELLATO ATTRAVERSO UNA MATRICE DI ADIACENZA CHE RAPPRESENTA COME

$$M_{Sij} = \begin{cases} \frac{1}{d_i} & \text{Se } i \rightarrow j \\ 0 & \text{ALTRIMENTI} \end{cases}$$

QUESTA MATRICE PRENDE IL NOME DI MATRICE DI ADIACENZA STOCASTICA CIÒ È LA SOMMA DI TUTTI I PESI SUCCESSIVI NELLA COLONNA È SEMPRE PARI A 1. AVENDO LA MATRICE IN QUESTA FORMA, LA SI PUÒ RISCRIVERE SOTTO FORMA DI ALGEBRA LINEARE COME:

$$z = M \cdot z$$

CHE PUÒ ESSERE ANCHE SCRITTA COME

$$1 \cdot z = M \cdot z \Rightarrow z = Mz$$

QUESTA FORMULAZIONE CI FA CAPIRE CHE z È L'VECTORE DELLA
MATRICE DI ADIACENZA M CON AUTOVALORI pari a 1. QUINN' SI
PUSCISSA TROVARE ATTRAVERSO LE FORMULE DELL'ALGEBRA LINEARE QUESTO
AUTOVETTORE z , VUOL DIRE CHE DEVO TROVARSI UN VETTORE CHE È MOLTIPLICATO
PER LA MATRICE DI ADIACENZA STOCHASTICA M DA SEMPRE z .

PER RISOLVERE QUESTO PROBLEMA POSSO UTILIZZARE IL METODO DELLE POTENZE
CIOÈ POSSO UTILIZZARE UN METODO ITERATIVO.

PARTENDO DA UNA CONFIGURAZIONE INIZIALE z_0 , DI LUNGHEZZA m ,
DOVE m È IL NUMERO DI PAGINI, CHE PUÒ ESSERE

$$z_0 = \begin{pmatrix} 1/m \\ 1/m \\ \vdots \\ \vdots \\ 1/m \end{pmatrix}$$

SUCCESSIONAMENTE z_1 SARÀ¹; $z_1 = M \cdot z_0$ E I SUCCESSIVI SARANNO:

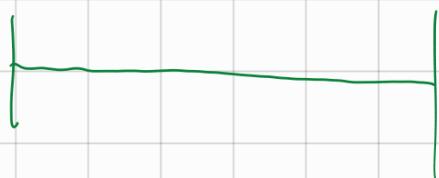
$$z_{t+1} = M z_t$$

POSSO TERMINARE IL PROCEDIMENTO DOPO UN NUMERO ARBITRARIO DI
ITERAZIONI OPPURE QUANDO LA DIFFERENZA TRA DUE VETTORI È
MINORE DI UN VALORE ϵ :

$$|r_{t+1} - r_t| < \epsilon$$

QUESTO PERCHÉ I VETTORI DOPO UN TOT DI PASSI SI STABILIZZANO AL PAGE RANK DI QUELLE PAGINE.

NELLA FORMULAZIONE BASATA SUL RANDOM SERVER L'IPOTESI E' CHE CI SIA UN NAVIGATORE CASUALE, CHE PARTE DA UNA PAGINA CASUALE E INIZIA A NAVIGARSI SEGUENDO I LINK. A LUNGO ANDARE IL PAGE RANK DELLE PAGINE CONSIDERA LA PROBABILITÀ CHE IL NAVIGATORE HA DI FINIRE NELLA PAGINA. I RISULTATI DEL SERVER NON VARIANO A SECONDA DEL NODO DAL QUALE INIZIANO.



DESCRIVERE I PROBLEMI DEL DEAD END E DEL SPIDER TRAP E ILLUSTRARNE UNA POSSIBILE SOLUZIONE

LO SPIDER TRAP SI HA QUANDO UN GRUPPO DI PAGINE SI LINKANO A VICENDA E DURANTE LA NAVIGAZIONE NON RIESCO A USCIRE DAL CERCHIO DI DIPENDENZA CHE SI E' CREATO.

IL DEAD END SI HA QUANDO ENTRI IN UNA PAGINA CHE NON HA LINK USCENTI E NE RIMANGI INTRAPPOLATO.

IL MECCANISMO PER USCIRE DA QUESTI PROBLEMI E' IL RANDOM TELEPORT. QUESTO MECCANISMO FA VARIARE LE EQUAZIONI DI FLUSSO



SIA Q UNA QUERY CHE HA 8 DOCUMENTI RILEVANTI NELLA COLLEZIONE,
 SUPPONIAMO CHE UN ALGORITMO DI RITROVAMENTO APPLICATO A Q RITORNI IL
 RANKING R_q : $D_3 D_1 D_5 D_7 D_8$. SUPPONIAMO CHE $D_1 \in D_7$ SIANO DOCUMENTI
 RILEVANTI PER Q. CALCOLARE PRECISION, RECALL, F1, R-PRECISION E AVERAGÉ
 PRECISION

D_3	D_1	D_5	D_7	D_8
0	X	0	X	0

$$AP = \frac{1/2 + 2/4}{8}$$

$$\text{PRECISION} = \frac{2}{5}$$

$$\text{RECALL} = \frac{2}{8}$$

$$F1 = \frac{2 \cdot P \cdot R}{P + R} = \frac{2 \cdot 2/5 \cdot 2/8}{2/5 + 2/8}$$

$$R\text{-PRECISION} = P @ 8 = 2/8$$

NUMERO DI DOCUMENTI RILEVANTI RICHIESTI
 WJM Doc. Rec.



$$C = \{c_1, c_2\} \quad V = \{T_1, T_2, T_3, T_4, T_5\} \quad K-NN (K=3)$$

$$TR = \{ \langle D_1, c_1 \rangle, \langle D_2, c_2 \rangle, \langle D_3, c_1 \rangle, \langle D_4, c_2 \rangle \}$$

$$d(T_1; 2, T_5; 2)$$

	T_1	T_2	T_3	T_4	T_5
D_1	3	3	0	4	0

\vec{d}_2 1 0 2 0 1

\vec{d}_3 0 1 0 2 1

\vec{d}_4 0 2 0 0 4

$$sim(\vec{d}, \vec{d}_1) = 2 \cdot 3 + 2 \cdot 0 = 6$$

$$sim(\vec{d}, \vec{d}_2) = 2 \cdot 1 + 2 \cdot 1 = 4$$

$$sim(\vec{d}, \vec{d}_3) = 2 \cdot 0 + 2 \cdot 1 = 2$$

$$sim(\vec{d}, \vec{d}_4) = 2 \cdot 0 + 2 \cdot 4 = 8$$

$\overbrace{\quad}$

\vec{d}_4 \vec{d}_1 \vec{d}_2 \vec{d}_3

c_2 c_1 c_2 c_1

.

NEL CLASSIFICATORE K-NN IL TRAINING SET È
IL CLASSIFICATORE.

CALCOLA LA SIMILARITÀ TRA \vec{d} E I VARI DOC.,
GLI ORDINA IN MANI



ESERCIZIO RILEVANCE FEEDBACK

SIA DATA LA SEGUENTE MATRICE TERMINI DOCUMENTI CONTENUTO

PESI TF-IDF NON NORMALIZZATI:

	T1	T2	T3	T4	T5	T6
D1	0,8	1,2	0	0,7	0	0
D2	0,1	1	1,4	0	0	0
D3	0	0,2	0	0	3,2	0,9
D4	0,1	0,1	0,1	0	2,3	1,7
D5	0	0	2	2	1	0

LA QUERY $q = (t_1:1, t_2:2)$

- 1) CALCOLARE IL RANKING DEI DOCUMENTI RISPETTO ALLA QUERY q UTILIZZANDO LA SIMILARITÀ DEL COSENZO
- 2) ASSUMENDO CHE IL TERZO È IL QUARTO DOCUMENTO DEL RANKING SIANO RILEVANTI E CHE IL PRIMO DOCUMENTO DEL RANKING NON SIA INVECE RILEVANTE, RIFORMULARE LA QUERY UTILIZZANDO L'ALGORITMO DI ROCCHIO E RECALCOLARE IL RANKING DEI DOCUMENTI

CALCOLO LA LUNGHEZZA DEI DOCUMENTI E DELLA QUERY

$$|d_1| = \sqrt{0,8^2 + 1,2^2 + 0,7^2} = 1,603$$

$$|d_2| = \sqrt{0,1^2 + 1^2 + 1,4^2} = 1,723$$

(1)

$$|d_3| = \sqrt{0,2^2 + 3,2^2 + 0,8^2} = 3,330$$

$$|d_4| = \sqrt{0,1^2 + 0,1^2 + 0,1^2 + 2,3^2 + 1,7^2} = 2,865$$

$$|d_5| = \sqrt{2^2 + 2^2 + 1^2} = 3$$

$$|q| = \sqrt{1^2 + 2^2} = 2,236$$

PESI NORMALIZATI %

$$T_1, D_1 = \frac{0,8}{1,603}$$

	t_1	t_2	t_3	t_4	t_5	t_6	
d_1	0,489	0,749	0	0,431	0	0	,
d_2	0,058	0,580	0,812	0	0	0	,
d_3	0	0,060	0	0	0,961	0,27	,
d_4	0,035	0,035	0,035	0	0,803	0,583	
d_5	0	0	0,667	0,667	0,333	0	
q	0,41	0,884	0	0	0	0	

CALCOLO DELLA SIMILARITÀ DEL COSENO

$$\cosim(d_1, q) = 0,958 \cdot 0,947 + 0,748 \cdot 0,834 = 0,883$$

$$\cosim(d_2, q) = 0,058 \cdot 0,947 + 0,580 \cdot 0,834 = 0,545$$

$$\cosim(d_3, q) = 0,060 \cdot 0,834 = 0,054$$

$$\cosim(d_4, q) = 0,035 \cdot 0,947 + 0,035 \cdot 0,834 = 0,047$$

$\cosim(d_5, q) = 0 \Rightarrow$ NON HANNO PAROLE IN COMUNE.

(2)

LA QUERY MODIFICATA È

$2 \cdot q_0 + \beta \cdot \text{CENTROIDE DOCUMENTI RILEVANTI} - \gamma \cdot \text{CENTROIDE DOCUMENTI N.R.}$

CALCOLO IL CENTROIDE PER I DOCUMENTI RILEVANTI.

RIPORTO I VALORI NORMALIZZATI

	T1	T2	T3	T4	T5	T6
$o(d_3)$	0	0,060	0	0	0,861	0,270
$o(d_4)$	0,035	0,035	0,035	0	0,803	0,583
CENTROIDE DOCUMENTI RILEVANTI	0,017	0,047	0,017	0	0,882	0,432

AVENDO SOLO 1 DOCUMENTO NON RILEVANTE, IL CENTROIDE
DEI DOCUMENTI NON RICEVANTI CORRISPONDE PROPRIO A SE STESSO

APPLICO IL METODO DI ROCCHIO

	T_1	T_2	T_3	T_4	T_5	T_6
q	0,447	0,844	0	0	0	0
CENTROIDE DOC RILEVANTI	0,018	0,067	0,014	0	0,882	0,432
CENTROIDE DOC NON RILEVANTI	0,489	0,743	0	0,434	0	0
$q_1 = q + 0,75 \cdot R_{RL} -$ $0,25 \cdot R_{NR}$	0,336	0,743	0,013	-0,109	0,661	0,324

I PESI NEGATIVI VENGONO ANNULLATI

$$q_1 \quad 0,336 \quad | \quad 0,743 \quad | \quad 0,013 \quad | \quad 0 \quad | \quad 0,661 \quad | \quad 0,324$$

CALCOLO NUOVARMENTE LA SIMILARITÀ DEL COSENZ

$$\cos(\alpha_1, q_1) = 0,675$$

$$\cos(\alpha_2, q_1) = 0,661$$

$$\cos(\alpha_3, q_1) = 0,767$$

$$\cos(\text{d}_4, q_1) = 0,761$$

$$\cos(\text{d}_5, q_1) = 0,156$$

d_3, d_4, d_1, d_2, d_5 .



$C = \{\text{SPORT}, \text{POLITICA}\}$

TRAINING SET	D1	SPORT
	D2	SPORT
	D3	SPORT
	D4	POLITICA
	D5	POLITICA

$$D_1 (3.8, 1, 0, 4.3, 5.8)$$

$$D_2 (0, 0, 1, 1, 4)$$

$$D_3 (0, 3.5, 1.4, 8, 6.3)$$

$$D_4 (1, 1, 0, 0, 0)$$

$$D_5 (5.5, 6.3, 0, 0.4, 0)$$

$$\vec{c}_{\text{SPORT}} = \underbrace{\vec{d}_1 + \vec{d}_2 + \vec{d}_3}_{3} =$$

$$\left(\frac{3.8}{3}, \frac{1+3.5}{3}, \frac{1+1.4}{3}, \frac{4.3+1+3}{3}, \frac{5.3+4+6.3}{3} \right)$$

$$c_{\text{POLITICA}} = \left(\frac{1+5.3}{2}, \frac{1+6.3}{2}, 0, \frac{0+0.4}{2}, 0 \right)$$

$$\vec{d} (1, 1, 1, 0, 1)$$

$$\text{COSIM}(\vec{d}, \vec{c}_{\text{SPORT}}) = 2$$

$$\text{COSIM}(\vec{d}, \vec{c}_{\text{POLITICA}}) = 0$$

SIANO DATI IN INPUT LA SEGUENTE QUERY $q \in$ IL DOCUMENTO D ,

ASSUMENDO CHE:

$q = \text{"INFORMATION RETRIEVAL"}$

$d = \text{"INFORMATION RETRIEVAL AND TEXT RETRIEVAL"}$

CALCOLARE LA SIMILARITÀ DEL COSENO TRA LA QUERY $q \in$ IL DOCUMENTO D , ASSUMENDO CHE:

- IL TERMINE AND SIA UNA STOPWORD

- IL DOCUMENTO FREQUENCY INFORMATION, RETRIEVAL E TEXT SIA RISPECTIVAMENTE 10, 50 E 100.
- IL NUMERO DEI DOCUMENTI NELLA COLLEZIONE SIA $N=100$
- SIA UTILIZZATO IL TF-IDF COME SCHEMA DI PESATURA DEI TERMINI NEL DOCUMENTO E NELLA QUERY

RAPPRESENTO Q E D SOTTOFORMA DI BAG OF WORDS

$$Q = \langle \text{INFORMATION}: 1, \text{RETRIEVAL}: 1 \rangle$$

$$D = \langle \text{INFORMATION}: 1, \text{RETRIEVAL}: 2, \text{TEXT}: 1 \rangle$$

IDF

$$\text{IDF}_{\text{INFORMATION}} = \log \frac{1000}{10} = 2$$

$$\text{IDF}_{\text{RETRIEVAL}} = \log_{10} \frac{1000}{50} = 1,3$$

$$\text{IDF}_{\text{TEXT}} = \log \frac{1000}{100} = \log 10 = 1$$

MATRICE TERMINI DOCUMENTI TF-IDF

	Q	D
INFORMATION	$1 \cdot 2 = 2$	$1 \cdot 2 = 2$
RETRIEVAL	$1,3 \cdot 1 = 1,3$	$1,3 \cdot 2 = 2,6$

TEXT

0

$$1 \cdot 1 = 1$$

$$\vec{a} (2, 1.3, 0) \quad \vec{b} (2, 2.6, 1)$$

$$\cos(\vec{a}, \vec{b}) \quad 2 \cdot 2 + 1.3 \cdot 2.6 + 0 = 0.87$$

$$\sqrt{2^2 + 1.3^2} \cdot \sqrt{2^2 + 2.6^2 + 1^2}$$

