

Social MINING

IIº SEM

2021



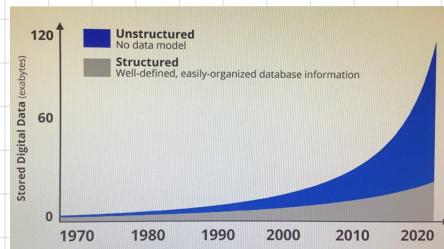
11/03

INFORMATION RETRIEVAL (IR)

IR - TASK

STRUCTURED vs
UNSTRUCTURED DATA

- Trovarci materiali (docs, video, images...)
- Spesso i dati sono **UNSTRUCTURED**, addirittura nelle grandi compagnie
 - ↳ **SEMI-STRUCTURED** sono ad esempio le slides con **FIELDS** come **TITLO** e **CONTENUTO**.
- **STRUCTURED**: Relational DBs, XML, ...



IR riguarda solo i 9
RETRIEVAL? → NO

- **CLOUDING**: cluster di documenti
- **CATEGORIZATION**: dato un insieme di topic assegnargli un documento
- **INFORMATION EXTRACTION**
- **QUESTION ANSWERING**: Rispondere a domande come: facts, How, Why
- **OPINION MINING**: analizzare il sentimento dietro i testi

Terminology

SEARCHING: ricerca di una **specific information**

BROWSING: **UNSTRUCTURED EXPLORATION**

CRAWLING: muoversi tra **HYPYERLINKS**

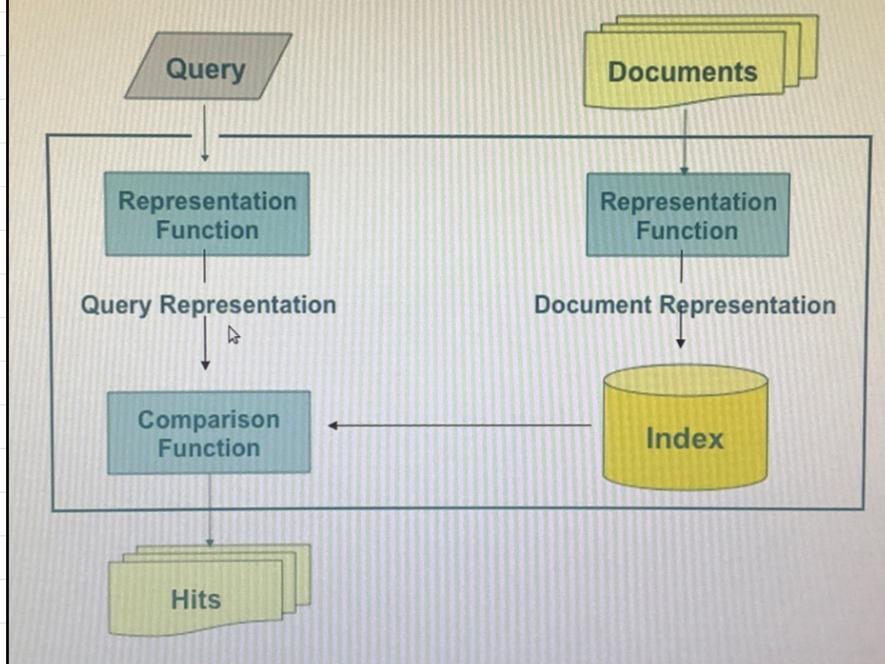
SCRAPING: estrarre contenuto da pagine

QUERY: **Strong**

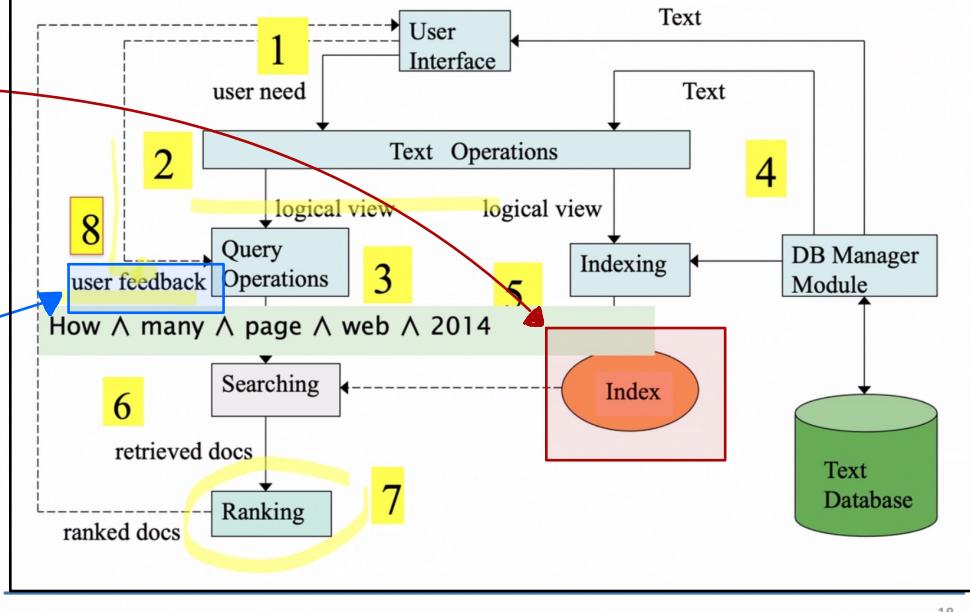
FULL-TEXT SEARCHING: comporre la query con ogni parola del testo

FIELDED SEARCH

Inside The IR Black Box



Workflow



L'INDEX è ciò che vogliamo creare al fine di indirizzare la query

Se l'utente dica un risultato, gli abbiamo dato ciò che cercava.

es. Può dipendere da FEEDBACK (oppure da PARTE RANK)

• Ottenuti i risultati come li presentiamo all'utente?

SORTING

RANKING by similarity tra query e documenti

RANKING by importance

DOCUMENT REPRESENTATION

. Dato un documento non strutturato \Rightarrow strutturato

(1) BAG OF WORDS MODEL : Sia un array dei Token

LDVARIANTI

HOW TO WEIGHT A WORD

\hookrightarrow es. BOOLEAN : Data una lista, assegna 1 se la Parola è nel documento

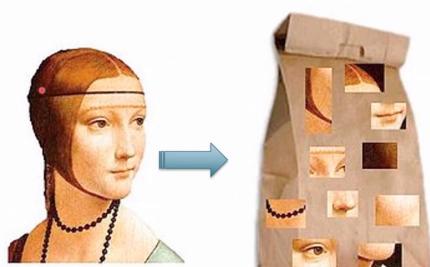
WHAT IS A WORD

\hookrightarrow SINGLE WORD

DOUBLE " : Nome e cognome

INCONDIZIONE : {go, gone, going...}, o
{polite, impolite, dormire, casa}

\hookrightarrow usato anche con immagini



DOCUMENT PARSENG

- composta scanning e trasformazione in bag of words
- ci servono 3 cose:
 - lingua (esiste MULTILINGUALITY)
 - tipo di file
 - character set
- {
 - è un singolo file?
 - è una singola pagina web o un sito?

UNIT DOCUMENT

Cos'è un
unità di
documento?

TOKENIZATION

ISSUES

- Un TOKEN è una sequenza di caratteri

- Come scegli i token?

- ↳ Prendo <Nome> - <sogno> insieme?
- ↳ Prendo punteggiatura?
- ↳ Prendo STOP WORDS?
- ↳ i numeri (DATE, NUM. DI TELEFONO, ...)
- ↳ Lingue (arabo, cinese, ...)

APPROCCI

STOP WORDS
• AND, A, TO,
↳ tempo stopgo?

NORMALIZATION

ACRONIMO

- U.S.A. => USA

SINONIMO

- CAR => AUTOMOBILE

TYPO

- GOOLGE => GOOGLE

CASE-FOLDING

- tutto in lowercase

LEMMATIZATION

- Riduzione a un LEMMA

es. {are, is, am} → be

{car, cars, car's} → car

STEMMING

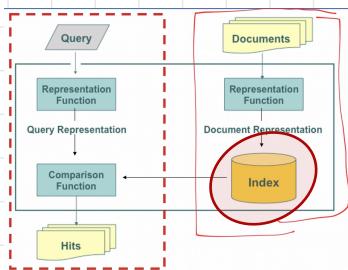
- Riduzione alla RADICE comune

for example compressed
and compression are both
accepted as equivalent to
compress.

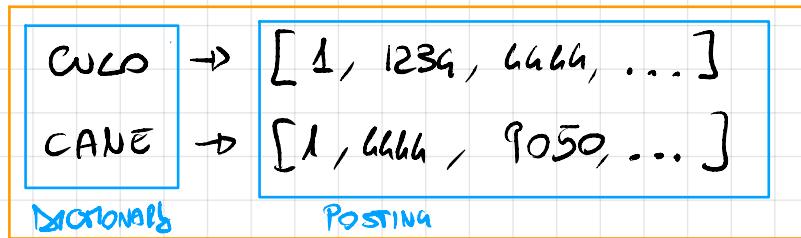
for example compress and
compress are both accepted
as equivalent to compress

17/03

INVERTED INDEX



- L'index è una lista che mappa ogni termine agli IDs dei documenti che lo contengono (POSTING LIST). È detto INVERTED perché non mappa DOC → TERMS MA TERM. → Docs.

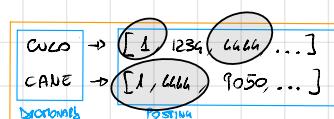


- Inoltre nel DICTIONARY non teniamo le occorrenze del termine totale, ma solo il numero di documenti in cui appare.

INDEX SEARCH

QUERY

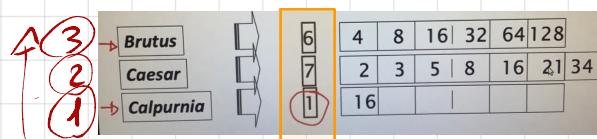
- query: "CUCO AND CANE"
→ Merghiamo i Postings di entrambi e riportiamo quindi i Docs {1, 6666}



OPTIMIZZAZIONE

- Merghiamo a partire da quello con POSTING più corto.

Lo ecco perché abbiamo tenuto i **DOC. & DOCS** in cui appare, così possiamo sapere a quali documenti appartiene.



PHRASE QUERY

- query: "RED BRICK HOUSE" → non va bene, il modo di prima, perché le parole sono legate

BI-WORD INDEXIS

- Rappresentiamo i due termini con dopp. termini
- Ad esempio dalla query: "STANFORD UNIVERSITY PAO ALTO" esce:

"STANFORD UNIVERSITY" AND "UNIVERSITY PAO" AND "PAO ALTO"

- Ovviamente la grandezza dell'INDEX cresce di molto

- Per ogni termine stiamo sia i documenti sia la posizione del termine nel documento, e teniamo anche la doc.freq.

<term, number of docs containing term;
doc1: position1, position2 ... ;
doc2: position1, position2 ... ;
etc.>

to, 993427: *
(1, 6: (7, 18, 33, 72, 86, 231);
2, 5: (1, 17, 74, 222, 255);
4, 5: (8, 16, 190, 429, 433);
5, 2: (363, 367);
7, 3: (13, 23, 191); ...)

be, 178239:
(1, 2: (17, 25);
4, 5: (17, 191, 291, 430, 434);
5, 3: (14, 19, 101); ...)

POSITIONAL INDEXES

Ad esempio per la query: "To BE or NOT To BE"

- Extract inverted index entries for each distinct term: **to, be, or, not.**
- Merge their *doc:position* lists to enumerate all positions with "to be or not to be".

• **to:**

- 2:1,17,74,222,551;
- 4:8,16,190,429,433;
- 7:13,23,191; ...

• **be:**

- 1:17,19;
- 4:17,191,291,430,434; 5:14,19,101; ...

① MELCO I DOCUMENTI
IN CW APPARE

② CONTROLLA SE LE
POSIZIONI IN UNO DOC.
SONO CONSECUTIVE

RELAXATION

- Possiamo rilassare la tecnica riferendoci a una certa distanza nei documenti

- For example: *employment /4 place*
Find all documents that contain EMPLOYMENT and PLACE within 4 words of each other.
- "*Employment agencies that place healthcare workers are seeing growth*" is a hit.
- "*Employment agencies that have learned to adapt now place healthcare workers*" is **not a hit.**

RANKING

BOOLEAN MODEL

- Query e Docs rappresentati come espressioni booleane



$$\begin{aligned} q = a \wedge (b \vee (\neg c)) &= \\ (a \wedge b \wedge c) \vee (a \wedge b \wedge (\neg c)) \vee (a \wedge (\neg b) \wedge (\neg c)) &\quad (\text{DNF form}) \\ \square \vee (q_{\text{dnf}}) &= (1,1,1) \quad (1,1,0) \quad (1,0,0) \end{aligned}$$

» Disjunctive Normal Form

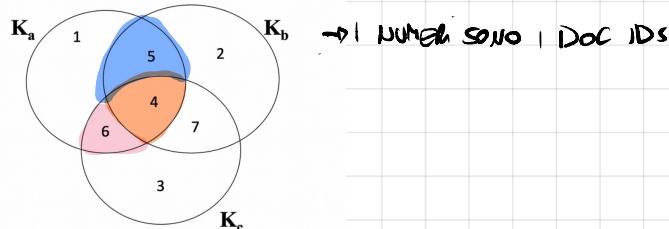
» Ex: $(apple, computer, red) \vee (apple, computer) \vee (apple)$

» $\square \vee (q_{\text{cc}}) = (1,1,0)$

» Conjunctive Component

c'è un MATH perché i documenti ci interessano

- Similar/Matching documents
- $md_1 = [apple apple blue day] \Rightarrow (1,0,0)$
- $md_2 = [apple computer red] \Rightarrow (1,1,1)$



$$\begin{aligned} q = k_a \wedge (k_b \vee k_c) \\ (1 \wedge 1 \wedge 1) \vee (1 \wedge 1 \wedge 0) \vee (1 \wedge 0 \wedge 1) \end{aligned}$$

Which one?

• Il problema è che tutti i documenti sono allo stesso livello,
e inoltre a una query possono essere restituiti Troppi / Pochi risultati
query piccola query grande

VECTOR MODELS
