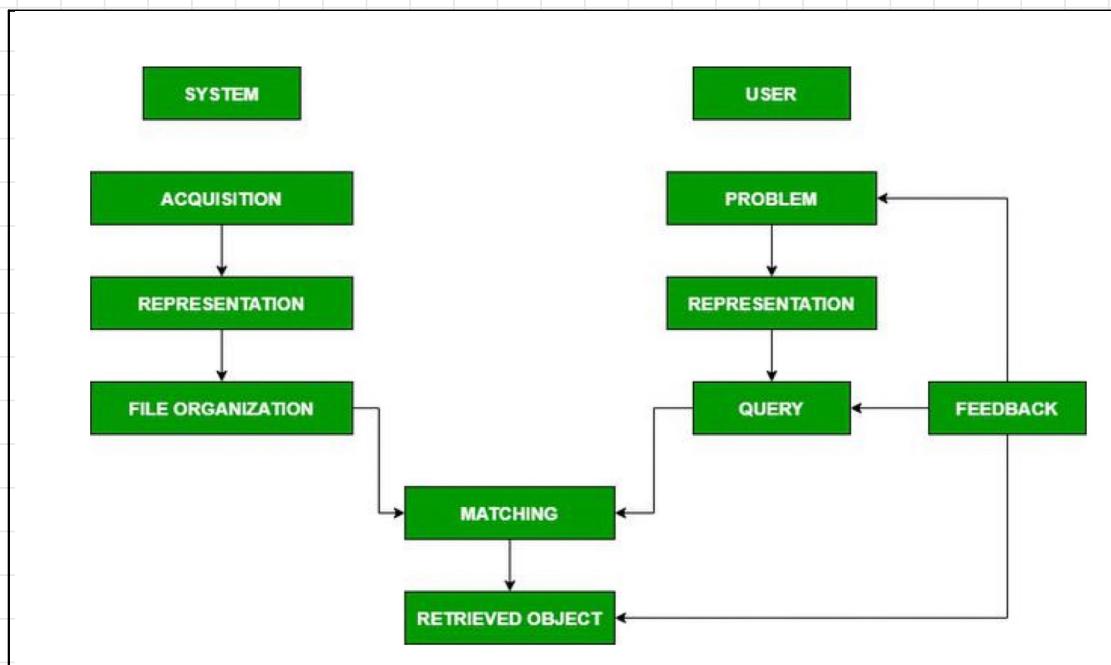


Information Retrieval

Un sistema di Information Retrieval (IR) è un insieme di tecnologie e algoritmi progettati per recuperare e presentare documenti non strutturati in modo rilevante agli utenti che cercano informazioni specifiche. Questi sistemi sono progettati per gestire documenti **non strutturati**, come testi, immagini o video. La struttura dei documenti può variare ampiamente, ma l'obiettivo è estrarre informazioni rilevanti da tali documenti.

Un sistema di IR è progettato per assistere gli utenti durante la fase di ricerca. Gli utenti formulano **query** o richieste di ricerca, e il sistema cerca di presentare loro risultati pertinenti in modo efficace. Il successo di un sistema di IR è valutato in base alla sua capacità di mostrare risultati pertinenti. Questo significa che i documenti presentati dovrebbero essere strettamente collegati alle intenzioni dell'utente e alle **parole chiave** specificate nella query.

IR Model



C'è differenza tra un'information retrieval e una semplice richiesta su un database.

Information Retrieval:

- Information Retrieval è un processo che coinvolge un programma software per l'organizzazione, l'archiviazione e la valutazione di informazioni da documenti o archivi di documenti, soprattutto informazioni testuali.
- Scopo principale: recuperare informazioni su un argomento specifico.
- È probabile che piccoli errori passino inosservati durante il recupero dei dati.
- Le query con cui opera non sono strutturate ed usano il linguaggio naturale che è spesso ambiguo.
- I risultati ottenuti sono dati anche da corrispondenze parziali tra la query e i documenti.
- I risultati sono solitamente ordinati in base alla rilevanza rispetto alla query.

Database Retrieval:

- Database Retrieval è un processo che mira a identificare e recuperare dati da un database in base a una query fornita dall'utente o da un'applicazione.
- Si basa sulla determinazione delle parole chiave nella query dell'utente per recuperare dati.
- Un singolo errore può significare un fallimento nel recupero dei dati.
- Le query con cui opera sono altamente strutturate, con una semantica ben definita per i dati.
- I risultati ottenuti sono esatti e rappresentano un accoppiamento preciso tra la query e i dati nel database.
- I risultati non sono ordinati per rilevanza, ma solitamente vengono restituiti nell'ordine in cui sono memorizzati nel database.

Text Retrieval

Gli elementi del Text Retrieval sono i seguenti :

- **Vocabolario:** Il vocabolario è l'insieme delle parole chiave o dei termini utilizzati nei documenti o nelle query.
- **Insieme delle Query:** L'insieme delle query è costituito dalle richieste degli utenti che vengono utilizzate per recuperare i documenti rilevanti dal sistema di IR.
- **Il Documento:** Un documento è un'unità informativa che può essere un articolo, un testo, una pagina web o qualsiasi altra fonte di informazioni.
- **Collezione di Documenti:** La collezione di documenti è l'insieme di tutti i documenti disponibili nel sistema di IR. La collezione costituisce la base dei dati da cui il sistema di IR recupera le informazioni.
- **Insieme dei Documenti Rilevanti:** Questo è l'insieme dei documenti che sono considerati rilevanti per una specifica query dell'utente. Si tratta di un sottoinsieme dell'insieme Collezione di Documenti. L'ordine di questi documenti può essere utilizzato per valutare l'attendibilità e l'efficacia del sistema di IR.

Modelli Text Retrieval

I modelli di text retrieval sono fondamentali nell'ambito dell'Information Retrieval (IR) per aiutare a organizzare e recuperare informazioni rilevanti da una vasta collezione di documenti in base alle query degli utenti.

• Modello Booleano

Il Modello Booleano rappresenta un approccio deterministico che fa uso di operatori logici per combinare i termini di una query e identificare i documenti che soddisfano le condizioni logiche definite.

Il Modello Booleano si basa sull'uso di operatori logici come "AND," "OR," e "NOT" per creare query complesse. Ad esempio, se desideriamo trovare documenti che contengono sia la parola "cat" che la parola "dog," possiamo utilizzare l'operatore "AND" scrivendo "cat AND dog" nella query. Questo significa che i documenti recuperati devono contenere entrambe le parole "cat" e "dog" per essere considerati rilevanti.

Funzionamento

A partire dalla collezione di documenti, viene creato un vettore di lunghezza fissa per ciascun documento, in cui ogni la posizione i -esima vale 0 o 1, in funzione della presenza o assenza di quella parola nel documento stesso. I vettori risultanti vengono quindi utilizzati per calcolare misure di similarità tra i documenti e le query degli utenti.

V = {I, like, cats, and, dogs, are, cute, fluffy, loyal, friendly, birds, colorful, sing, beautifully}

D1: I like cats and dogs.

[1 1 1 1 1 0 0 0 0 0 0 0 0]

Si calcola il risultato logico della query sostituendo 0 se la parola nella query non è presente nel vocabolario o 1 se è presente. Si valuta poi il risultato dell'espressione logica ottenuta.

Viene creato un vettore a partire dalla query sostituendo, per ogni termine nel vocabolario, 0 se il termine non figura nella query, oppure il risultato logico della query calcolato in precedenza.

Q = (cat AND dog) OR bird → **Q = [0 0 1 0 1 0 0 0 0 1 0 0 0]**

[0 0 (1 AND 1) OR 1 0 (1 AND 1) OR 1 ...]

V = {I, like, cats, and, dogs, are, cute, fluffy, loyal, friendly, birds, colorful, sing, beautifully}

Si valuta infine una misura di similarità di ogni documento con la query facendo la somma dei prodotti di ogni bit tra il vettore dell' i -esimo documento e il vettore della query. I documenti che avranno ottenuto la misurazione di similarità più alta saranno rappresentativi dei risultati più pertinenti.

S (D1, Q) = [(0*1) + (0*1) + (1*1) + ...] = [2]

D1: [1 1 1 1 1 0 0 0 0 0 0 0 0] [2]
D2: [0 0 1 1 0 1 1 1 0 0 0 0 0] [1]
D3: [0 0 0 1 1 1 0 0 1 1 0 0 0] [1]

Limiti

- **Sono molto rigidi e difficili da esprimere per richieste complesse dell'utente:** modelli booleani sono noti per la loro rigidità e possono risultare complessi da utilizzare per esprimere richieste complesse da parte degli utenti. Questa rigidità può limitare la capacità del sistema di adattarsi alle preferenze e ai requisiti informativi più dettagliati degli utenti.
- **È difficile controllare il numero e la qualità dei documenti recuperati:** gli utenti possono trovare difficile ottenere il giusto equilibrio tra precisione e copertura, il che può portare a un eccesso di risultati o a un insieme insufficiente di documenti rilevanti. La modulazione del processo di recupero per ottenere una combinazione ottimale di documenti può risultare complicata.
- **Sono difficili da eseguire per ottenere un feedback di rilevanza:** la natura binaria del recupero booleano (documenti considerati rilevanti o non rilevanti) rende difficile l'integrazione efficace del feedback degli utenti, a differenza di modelli come quelli probabilistici o basati su spazio vettoriale, dove il feedback sulla rilevanza può essere integrato più agevolmente.

Vector Space Model

La rappresentazione dei documenti e delle query avviene come vettori nello spazio Euclideo multidimensionale. La similarità tra i documenti e la query viene calcolata con la similarità del coseno dell'angolo formato tra i vettori.

Le principali differenze col modello precedente sono le seguenti :

Corrispondenza parziale:

- Modello dello Spazio Vettoriale: la corrispondenza tra una query e i documenti non è binaria ma graduale e viene misurata in base alla similarità tra la query e i documenti. Questo consente di trovare documenti che corrispondono parzialmente alla query, assegnando loro punteggi di similarità.
- Modello Booleano: Nel Modello Booleano, la corrispondenza è binaria. Un documento è considerato rilevante solo se soddisfa tutti i termini specificati nella query. Non c'è una misura graduale di rilevanza. Questo significa che i documenti corrispondenti parzialmente alla query vengono trattati alla stessa stregua di quelli che soddisfano completamente la query.

Ranking:

- Modello dello Spazio Vettoriale: utilizza una classificazione graduale dei documenti in base alla loro rilevanza rispetto alla query. Gli utenti possono quindi ricevere una lista ordinata di documenti in base alla loro rilevanza stimata.
- Modello Booleano: la classificazione è binaria. Un documento è classificato come "rilevante" o "non rilevante" rispetto a una query specifica. Non c'è una graduale gerarchia di rilevanza. I documenti rilevanti sono restituiti senza ulteriori distinzioni.

Distanza tra due vettori

Possiamo calcolare la **L2-norm** della differenza tra i due vettori per fornire una misura del grado di separazione tra i vettori. Questa distanza viene utilizzata per calcolare la similarità tra la rappresentazione vettoriale di una query e quella di un documento. Maggiore è la distanza, minore è la similarità tra i due vettori.

Pertanto, la L2-norm della differenza può essere utilizzata per classificare i documenti in base alla loro rilevanza rispetto a una query, con documenti più simili alla query che otterranno una distanza euclidea inferiore.

$$\text{L2-norm} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2}$$

Similarità del coseno

La similarità del coseno è una metrica che misura il coseno dell'angolo tra due vettori proiettati in uno spazio multidimensionale. In altre parole, misura quanto simili sono due vettori considerando l'angolo tra di essi nello spazio vettoriale. Questo valore viene utilizzato per determinare quanto un documento è simile a una query.

$$\cos \theta = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|_2 \|\mathbf{b}\|_2}$$

Calcoliamo quindi la similarità del coseno come il rapporto tra il prodotto dei due vettori ed il prodotto delle loro norme Euclidee. I possibili valori della similarità del coseno rappresentano il grado di somiglianza o dissomiglianza tra due vettori in uno spazio vettoriale:

- Se vale **-1** indica che i vettori sono fortemente opposti. Questo significa che non c'è nessuna sovrapposizione o correlazione tra la query ed il documento.
- Se vale **0** indica che i vettori sono ortogonali. Ciò significa che non c'è sovrapposizione significativa tra di essi e di fatto sono indipendenti.
- Se vale **1** indica che i vettori si sovrappongono completamente. Significa che i vettori sono identici o perfettamente sovrapposti. La similarità è massima, indicando una forte somiglianza tra i vettori.

Distanza del coseno

Una misura più diretta è invece data dalla distanza del coseno, ottenuta come **1 - similarità**. Questa misura risulta più diretta poiché quantifica direttamente la differenza tra i documenti e la query, ad esempio :

- Se vale 0 indica che la similarità tra i vettori è massima
- Se vale 1 indica che non c'è sovrapposizione significativa tra i vettori

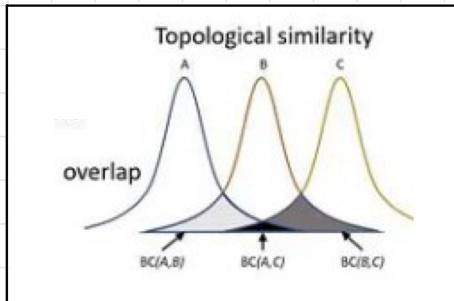
Distanza di Bhattacharyya

La distanza di Bhattacharyya è una misura di similarità tra due distribuzioni di probabilità ed è spesso utilizzata in contesti come il recupero di immagini o il riconoscimento di modelli. Questa misura fornisce informazioni sulla quantità di sovrapposizione tra le due distribuzioni, il che aiuta a determinare quanto siano simili o dissimili tra loro.

La formula per calcolare la distanza di Bhattacharyya tra due distribuzioni di probabilità P e Q è la seguente:

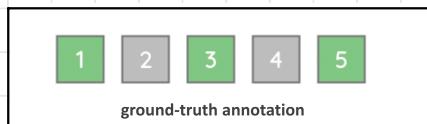
$$BD(P,Q) = -\ln(BC(P,Q))$$

Si valuta la similarità sulla base di quanto è grande l'area di overlap tra le distribuzioni in esame



Misure di Valutazione

Come valuto le performance di un algoritmo di Information Retrieval ? In base alla mia query il modello produce dei documenti. Si annotano i documenti con delle valutazioni di ground-truth , ovvero assegno un etichetta booleana di rilevanza ai documenti prodotti.



Bisogna chiarire il significato delle varie annotazioni. Supponendo di avere un sistema di rilevazione di immagini e di ricercare “Hot-dog” :

- vero positivo : le immagini appartengono alla classe Hot-dog ed il sistema le ha rilevate come tali.
- falso positivo : in realtà le immagini non appartengono alla classe Hot-dog ma il sistema le rileva come appartenenti.
- vero negativo : le immagini non appartengono alla classe cercata ed il sistema le rivela come tali.
- falso negativo : all’atto pratico l’immagine appartiene alla classe ma il modello non lo ha rilevato.

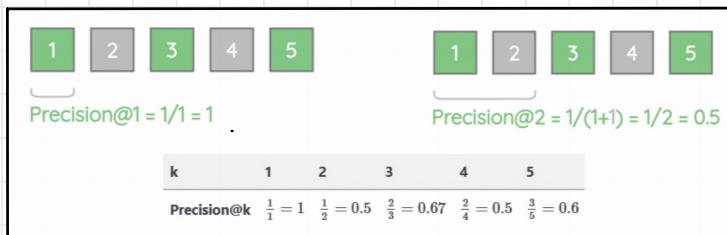
Possiamo suddividere le metriche di valutazione in posizionali e non posizionali a seconda del fatto che tengono conto della posizione dei risultati rilevanti.

- **Precision** (non posizionale) :

La Precision è una delle metriche chiave utilizzate per valutare quanto sia preciso un sistema di Information Retrieval (IR) nel fornire risultati rilevanti agli utenti.

$$\text{Precision}@k = \frac{\text{true positives}@k}{(\text{true positives}@k) + (\text{false positives}@k)}$$

La Precision restituirà un valore compreso tra 0 e 1, in cui un valore più vicino a 1 indica che il sistema di IR è più preciso nel recuperare i documenti rilevanti tra i primi k risultati.



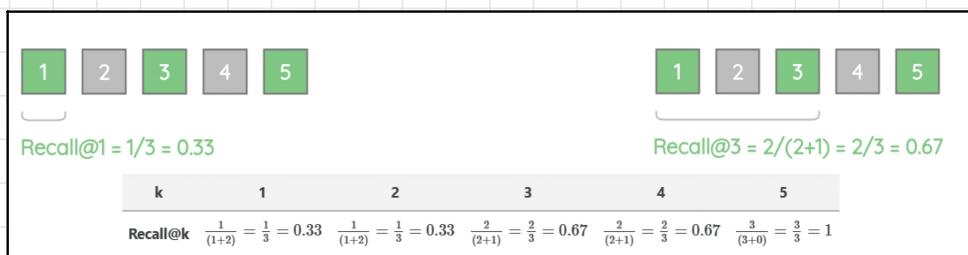
Tuttavia, la Precision ha una limitazione significativa: essa non tiene conto della posizione specifica dei documenti rilevanti all'interno della lista dei risultati. Ogni documento rilevante è considerato allo stesso modo, indipendentemente dalla sua posizione.

- **Recall** (non posizionale)

Il suo obiettivo principale è indicare quanti dei documenti rilevanti effettivamente presenti sono stati identificati dal sistema. In altre parole, il Recall misura la capacità del sistema di Information Retrieval di catturare tutti i risultati rilevanti tra quelli effettivamente disponibili.

$$\text{Recall}@k = \frac{\text{true positives}@k}{(\text{true positives}@k) + (\text{false negatives}@k)}$$

Il valore di Recall è sempre compreso tra 0 e 1, dove un valore più vicino a 1 indica che il sistema è in grado di recuperare una maggiore percentuale di documenti rilevanti disponibili per la query. In altre parole, un alto valore di Recall indica una maggiore capacità del sistema di non trascurare i documenti rilevanti.



- **F1-Score** (non posizionale)

L'F1-score è una metrica che rappresenta una combinazione delle metriche Precision e Recall. È particolarmente utile quando si desidera ottenere una valutazione complessiva delle prestazioni di un sistema di IR, tenendo conto sia della precisione che della capacità di recuperare tutti i risultati rilevanti.

$$F1@k = \frac{2 * (Precision@k) * (Recall@k)}{(Precision@k) + (Recall@k)}$$

- **Mean Reciprocal Rank (MRR)** (posizionale):

Il Mean Reciprocal Rank (MRR) è una metrica utilizzata per valutare l'efficacia di un sistema di Information Retrieval con particolare enfasi sul posizionamento del primo risultato rilevante restituito dal sistema. È utile quando si desidera che il sistema restituisca il miglior risultato rilevante e che questo risultato sia posizionato più in alto possibile nella lista dei risultati.

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

totale di query

MRR calcola l'inverso della posizione (rango) del primo risultato rilevante. Questo significa che, se il primo risultato rilevante è al primo posto, l'inverso della sua posizione sarà 1. Se è al secondo posto, l'inverso della posizione sarà 1/2, e così via.

MRR valuta l'efficacia del sistema nel restituire il risultato più rilevante quanto prima nella lista dei risultati, ma non considera quanti altri risultati rilevanti possano seguire.

Reciprocal Rank						
Query 1	1	2	3	4	5	$1/1 = 1$
Query 2	1	2	3	4	5	$1/2 = 0.5$
Query 3	1	2	3	4	5	$1/5 = 0.2$
						$MRR = (1+0.5+0.2)/3 = 0.567$

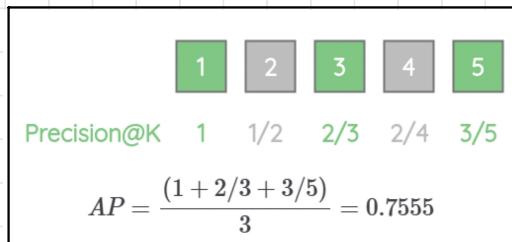
- **Average precision** (posizionale)

L'AP misura la precisione media in termini di posizionamento dei documenti rilevanti. La metrica considera la precisione a ogni posizione in cui un documento rilevante è trovato e quindi calcola la media di queste precisioni.

Questo significa che l'AP tiene conto sia del fatto che i documenti rilevanti sono stati effettivamente identificati che della loro posizione specifica nella lista dei risultati.

$$AP = \frac{\sum_{k=1}^n (P(k) * rel(k))}{\text{number of relevant items}}$$

L'AP restituirà un valore compreso tra 0 e 1, dove un valore più vicino a 1 indica che il sistema ha posizionato i documenti rilevanti in alto nella lista dei risultati in modo coerente. In altre parole, un alto AP indica una maggiore capacità del sistema di posizionare correttamente i documenti rilevanti.



- **Mean average precision** (posizionale)

La average precision viene calcolata a partire da una singola query. Per valutare l'efficienza di un sistema sarebbe opportuno valutare tale metrica su diverse query. La mean average precision (MAP) esegue una semplice media della AP calcolate sulle varie query.

Per ciascuna query, si calcola l'Average Precision, che rappresenta la precisione media per quella query specifica. Dopo aver calcolato l'AP per tutte le query, si procede a calcolare la media di questi valori AP.

$$MAP = \frac{1}{Q} \sum_{q=1}^Q AP(q)$$

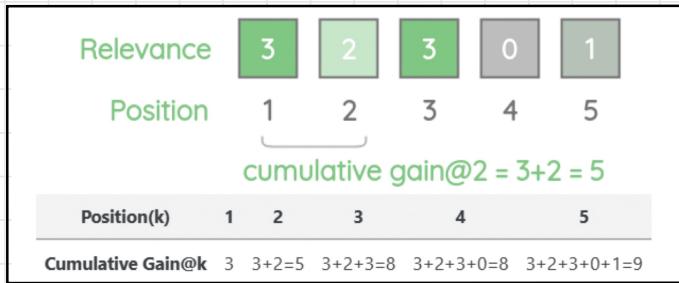
Potrebbe essere utile assegnare un **grado di rilevanza** al documento e non un valore booleano.



- **Cumulative Gains** (non posizionale)

La metrica dei Cumulative Gain è una misura che si basa su un'idea semplice: sommare progressivamente i valori di rilevanza dei primi k documenti man mano che si scorre la lista dei risultati.

$$CG@k = \sum_{1}^k rel_i$$



Questa metrica non distingue tra la posizione dei documenti rilevanti, ma si concentra esclusivamente sulla somma progressiva dei valori di rilevanza dei documenti nella lista dei risultati. Questo significa che, indipendentemente dalla posizione in cui si trovano i documenti rilevanti, il Cumulative Gain valuta semplicemente la somma cumulativa delle loro rilevanze. Pertanto, non fornisce informazioni sulla capacità del sistema di posizionare i documenti rilevanti in posizioni più alte o più basse nella lista dei risultati.

- **Discounted cumulative gains** (posizionale)

Questa metrica introduce una funzione di penalità basata su logaritmo per ridurre il punteggio di rilevanza in ciascuna posizione. In sostanza, il DCG tiene conto non solo della rilevanza dei risultati, ma anche della posizione in cui appaiono nella lista dei risultati.

Si calcola il DCG sommando la rilevanza di ciascun risultato normalizzata dalla funzione di penalità basata su logaritmo. Questa penalità attribuisce un peso inferiore ai risultati che compaiono più in basso nella lista.

$$DCG@k = \sum_{i=1}^k \frac{rel_i}{\log_2(i+1)}$$

↓
indice del documento

- **Normalized Discounted Cumulative Gain** (posizionale)

La Normalized Discounted Cumulative Gain (nDCG) è una metrica che tiene conto sia della rilevanza dei risultati che della loro posizione nella lista dei risultati, fornendo una valutazione normalizzata delle prestazioni di un sistema di Information Retrieval.

Si calcola la DCG ideale mettendo i documenti rilevanti ad inizio ordinati (per rilevanza decrescente).

Relevance	3	2	3	0	1
Position	1	2	3	4	5
Ideal Order of Items					
Relevance	3	3	2	1	0
Position	1	2	3	4	5

Si prosegue calcolando il rapporto tra la DCG e la IDCG(quella ideale) ottenendo un valore normalizzato e quindi confrontabile con i valori ottenuti da altre query.

$$NDCG@k = \frac{DCG@k}{IDCG@k}$$