9. Information Retrieval Evaluation

È importante ricordare che il **bisogno informativo dell'utente** viene espresso attraverso una **query**, ma la rilevanza deve essere valutata rispetto al **bisogno sottostante**, non alla query in sé. Ad esempio:

Bisogno informativo: *Il fondo della mia piscina sta diventando nero e necessita di essere pulito*
Query: *pool cleaner*

In questo caso, è necessario valutare se un documento risponde al bisogno reale dell'utente, non semplicemente se contiene le parole della query.

Giudizi di rilevanza

Nel caso più semplice, i giudizi di rilevanza sono **binari** (rilevante o non rilevante). Tuttavia, in alcuni scenari, possono essere espressi in modo più preciso con scale ordinali (ad esempio: 0, 1, 2, 3, ...).

Se per ogni query si considerasse l'intero insieme di documenti da giudicare, la valutazione della rilevanza risulterebbe estremamente onerosa e costosa. Per affrontare questo problema si utilizza la tecnica della **depth-k pooling**.

Depth-k pooling

Si prendono in considerazione i primi k documenti (ad esempio 100) restituiti da N sistemi di recupero dell'informazione differenti (ad esempio 100). Gli esseri umani devono quindi giudicare un "pool" di massimo $k \times N$ documenti (ad esempio 10.000), un numero significativamente inferiore rispetto all'intera collezione documentale, che potrebbe contenere milioni di documenti.

Misure di efficacia

Per valutare l'efficacia di un sistema di recupero dell'informazione (cioè la qualità dei risultati restituiti per una query), si utilizzano due parametri fondamentali relativi ai risultati ottenuti:

Precisione (Precision): rappresenta la frazione dei documenti restituiti che sono effettivamente rilevanti per il bisogno informativo dell'utente.

$$\text{Precision} = \frac{\text{documenti rilevanti recuperati}}{\text{documenti recuperati}} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Richiamo (Recall): rappresenta la frazione dei documenti rilevanti presenti nella collezione che sono stati effettivamente recuperati dal sistema.

$$\text{Recall} = \frac{\text{documenti rilevanti recuperati}}{\text{documenti rilevanti totali nella collezione}} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Compromesso tra precisione e richiamo

In uno scenario ideale, in cui i dati sono perfettamente separabili, sia la **precisione** sia il **richiamo** possono raggiungere il valore massimo di 1. Purtroppo, nella maggior parte delle situazioni pratiche, non è possibile ottenere contemporaneamente una precisione e un richiamo elevati. Aumentare la precisione porta spesso a una diminuzione del richiamo, e viceversa.

F-Measure

Presi singolarmente, né la precisione né il richiamo offrono una visione completa. È possibile avere un'ottima precisione ma un richiamo pessimo, oppure un ottimo richiamo ma una precisione scarsa. La **F-Measure** combina entrambi in un'unica misura, fornendo una valutazione complessiva dell'accuratezza del sistema di recupero dell'informazione.

La F-Measure tradizionale è calcolata come segue:

$$F\text{-}Measure = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Questa è la **media armonica**, nota anche come **F-Score** o **F1-Score**. Esiste una versione più generale, il **Fβ-score**, che applica pesi differenti (β) per valorizzare maggiormente la precisione oppure il richiamo, a seconda delle esigenze specifiche.

Valutazione dei risultati ordinati

Le misure di precisione, richiamo e F-Measure sono basate su insiemi **non ordinati** di documenti. Tuttavia, per valutare i risultati ordinati, come quelli forniti dai motori di ricerca moderni, è necessario estendere queste metriche.

Nei sistemi di recupero ordinato, l'insieme naturale di documenti recuperati da valutare è costituito dai primi k documenti restituiti. Di conseguenza, si introducono misure di **precisione e richiamo calcolate su un numero fisso k di documenti**, dette:

Precision @ K: frazione di documenti rilevanti tra i primi k risultati.
Recall @ K: frazione dei documenti rilevanti totali che compaiono tra i primi k risultati.

Queste misure sono spesso indicate anche come **P@K** e **R@K** e sono fondamentali per valutare l'efficacia delle ricerche nelle prime posizioni, ovvero quelle che l'utente vede subito nella pagina dei risultati.

Precisione @ K

Per calcolare la **Precisione @ K**, si imposta una soglia di rango K, corrispondente al numero di documenti recuperati, e si calcola la proporzione dei documenti rilevanti tra i primi

K restituiti. Tutti i documenti con un rango inferiore a K vengono ignorati.

Richiamo @ K

Anche per il **Richiamo @ K**, si imposta una soglia K e si calcola la proporzione di documenti rilevanti trovati tra i primi K risultati. Come nel caso della precisione, i documenti classificati oltre la soglia K non vengono considerati.

Average Precision (AP)

Analogamente alla F-Measure, l'**Average Precision** (Precisione Media) nasce dall'esigenza di utilizzare una misura aggregata, soprattutto quando il valore di K varia.

L'Average Precision è una misura aggregata per i risultati ordinati. Il suo calcolo avviene nel seguente modo:

Invece di fissare arbitrariamente un valore di K, si continua a considerare i risultati fino a quando **tutti i documenti rilevanti** sono stati recuperati, cioè fino al valore di K per cui **Recall @ K = 1**.

Si calcolano le **Precision @ K** solo per quei valori di K in cui è stato effettivamente recuperato un documento rilevante. La media di tutte queste misure di precisione costituisce l'**Average Precision (AP)**.

Mean Average Precision (MAP)

Quando si valuta un sistema di recupero, solitamente si misura l'efficacia su più query (ad esempio, 10.000 nella collezione TREC GOV2). Il numero di query rappresenta una dimensione aggiuntiva di aggregazione per valutare il sistema IR.

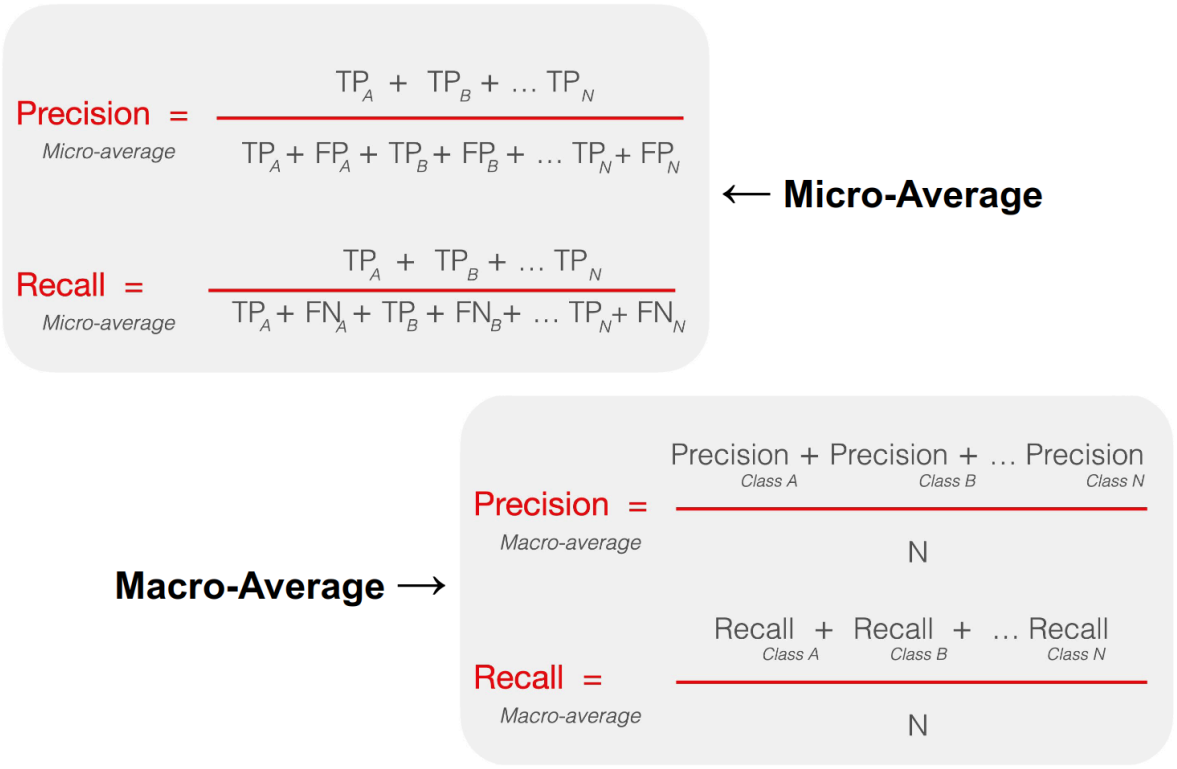
Dopo aver calcolato la **Average Precision** per ciascuna query nella collezione di test, si ottiene la **Mean Average Precision (MAP)** facendo la media delle AP su tutte le query.

Osservazioni sul MAP

Il **MAP** presuppone che l'utente sia interessato a trovare **molti documenti rilevanti** per ciascuna query. Se un documento rilevante non viene mai recuperato, si assume che la precisione relativa a quel documento sia zero. Il MAP rappresenta una **media macro**: ogni query contribuisce in egual misura al punteggio finale.

In genere, esiste maggiore concordanza nei valori di MAP per un determinato bisogno informativo tra diversi sistemi, rispetto ai valori di MAP ottenuti per bisogni informativi diversi usando lo stesso sistema.

Micro-media vs macro-media



Oltre la rilevanza binaria

Fino a ora si è assunta una nozione **binaria** di rilevanza: un documento è o **rilevante** oppure **non rilevante** rispetto a una query. Tuttavia, alcuni documenti possono essere **meno rilevanti di altri**, pur essendo comunque rilevanti. Questa visione **non binaria** richiede metriche specifiche per la valutazione.

Misure come **DCG (Discounted Cumulative Gain)** o **NDCG (Normalized Discounted Cumulative Gain)** sono progettate per tenere conto di differenti gradi di rilevanza. Nonostante ciò, l'uso della rilevanza binaria rimane più comune e continua a offrire una buona stima per la valutazione dei sistemi di recupero dell'informazione.