Performance Evaluation in (CS) IR System

Performance Evaluation in IR is the process of measuring how well an information retrieval system retrieves relevant documents for a user's query.

It helps us Compare systems, improve algorithms, and ensure users get accurate results.

Purpose of IR Evaluation System

- ✓ To check accuracy of retrieved results.
- ✓ To compare different retrieval models.
- ✓ To improve ranking algorithms
- ✓ To assess User Satisfaction

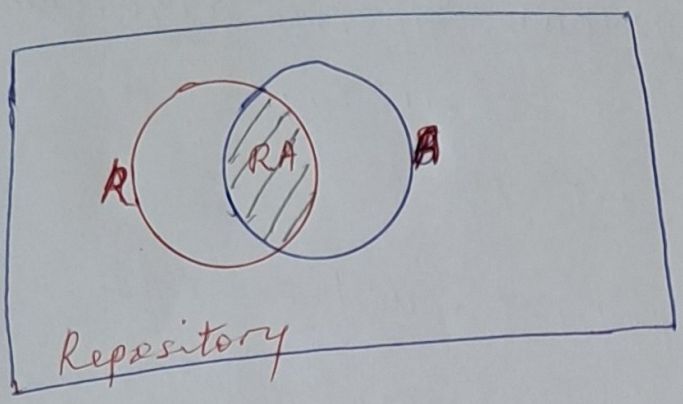
Key Concepts

* Relevant documents
  - documents that actually match the User's info need.
* Retrieved documents
  - documents returned by system (whether relevant OR NOT)

True/False classification

Category	Meaning
True positive (TP)	✓ Relevant documents retrieved
False positive (FP)	✓ NOT-relevant retrieved
False Negative (FN)	✓ Relevant but NOT retrieved
True Negative (TN)	✓ Non-relevant and not retrieved (Usually ignored in IR)

	Relevant	NOT-Relevant
Retrieved	TP	FP
NOT-Retrieved	FN	TN



RA = Retrieved Relevant

R = Relevant

A = Retrieved

Core Evaluation Metrics

precision :- measures accuracy :- of all retrieved documents, how many are Relevant?

$$\text{precision} = \frac{RA}{A} = \frac{TP}{TP+FP}$$

High precision => More relevant results at the top.

Recall :- Measure Completeness :- of all Relevant documents, how many were retrieved?

$$\text{Recall} = \frac{RA}{R} = \frac{TP}{TP+FN}$$

High Recall → System retrieves most of the relevant items.

F1-Score :- Harmonic Mean of precision and Recall.

$$\text{F1-score} = \frac{2\,(\text{precision} \times \text{Recall})}{\text{precision} + \text{Recall}}$$

Performance Evaluation in IR System is commonly Measured in terms of :-

Efficiency and Effectiveness.

Efficiency :- "Doing things the right way" refers how well the System uses time, Memory and Computation to produce results.

↳ Fast Searching (low response time)

- The System returns result quickly, even for large collections.
- Efficient indexing and query processing techniques are used.

Example: Using an inverted index instead of Scanning all documents

↳ High throughput

- The System can handle Many queries per second
- Efficient algorithms allow many users to search at once.

Example: Google processing Millions of queries Simultaneously

↳ Efficient use of Memory & Storage.

- Index structures are Compact and optimized
- Unnecessary data is not Stored

Example: Using Compression for posting lists

Scalable Architecture

- The system grows well with more data & More users
- performance doesn't collapse when the Collection size increases.

Example: Distributed indexing in search engines

Minimizing Computational cost

- Algorithms avoid heavy operations
- Query processing is optimized

Example: Using skip pointers to avoid scanning long posting lists

Summary:

* "Doing things the right way" in an IR system means retrieving information efficiently — fast, scalable, and resource friendly — without wasting time, memory, or computing power

Effectiveness = 'Doing the right things'

Effectiveness measures how well the system achieves its intended goals

In IR:
- Does the system returns relevant documents?
- How accurate are the result?
- Common metrics
  Precision, Recall, F1-score & Accuracy.

* Effectiveness focuses on quality & accuracy

Precision = the number of Relevant Retrieved documents / Total number of Retrieved documents.

Interpretation
"of all documents the system Retrieved, how many were actually Relevant?"

High precision is the system Retrieves mostly Relevant documents

low precision is the system Retrieves many IRRelevant documents (noise)

Example Suppose a system Retrieves 10 documents for a query.

Relevant Retrieved = 7
IRRelevant Retrieved - 3

So precision = $\frac{7}{10} = 0.7 = 70\%$.

* precision is very important when users need very accurate Results.

Recall is a core effectiveness metrics used to measure how well an IR system Retrieves all Relevant documents.
It is the proportion of Relevant documents that the system Succesfully Retrieved.

$$\text{Recall} = \frac{\text{Relevant Retrieved documents}}{\text{Total relevant documents.}}$$

Interpretation of all the Relevant documents that exist how many did the system find.

High Recall: System retrieves most of the Relevant documents.

low Recall: System missed many relevant documents.

Example: Suppose in the entire db
- There are 20 Relevant documents
- The system retrieves 12 Relevant documents.

$$\text{Recall} = \frac{12}{20} = 0.6 = 60\%.$$

- Recall is more important when:
  - Missing a relevant document is dangerous or costly.

F1-score also called F1-Measure is a metric used to Evaluate IR system by Combining precision and Recall into a single score.
It is useful when both precision and Recall are important and must be balanced.
It is the harmonic mean of precision and Recall.

$$F1\text{-score} = \frac{2\,(\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

* It gives Higher score when both precision and Recall are high.

Accuracy in IR system :- This is a metric that measures how often an IR system correctly classifies documents as Relevant or non-Relevant.
It considers all four outcomes :-

$$\text{Accuracy}(A) = \frac{TP + TN}{TP + TN + FP + FN}$$

Example:
Suppose an IR system processes 100 documents
TP = 30 (Relevant retrieved) TN = 60 (Not Relevant Not Retrieved)
FP = 5 FN = 5

$$A = \frac{30+60}{30+60+5+5} = \frac{90}{100} = 0.9 = 90\%$$

The notion of Relevance in IR:

* Relevance is a central concept in information retrieval (IR) that measures how well a retrieved document meets the user's information need. A document is considered relevant if it satisfies the user's query intention, provides useful information, or helps solve the user's problem. Relevance is subjective, meaning different users may judge the same document differently depending on their knowledge, background, purpose and context.

Relevance in IR is not just about matching keywords, it reflects how well the retrieved information aligns with the user's real info need, which may not always be full expressed in the initial query. Because of this, IR systems often incorporate methods like Relevance feedback, Query expansion, and Ranking algorithms to estimate and improve relevance.

Relevance can be understood in different forms, such as topical relevance (content matches the topic), cognitive relevance (fits the user's understanding), situational relevance (supports the users task), and contextual relevance (useful within a specific scenario).

Measuring relevance is essential for evaluating IR system performance through metrics such as Precision, Recall, and F1-score measure.

⇒ Relevance is a subjective judgment and may include:-
- ✓ Being timely (recent information)
- ✓ Being authoritative (from a trusted source)
- ✓ Satisfying the goals of the user and his/her intended use of the information (information need).

- Relevant information is that suited to your info need.
  What is actually needed (relevant)
- Dependent on: (User, space/time, Group and context)

IR is very concerned with Relevance.

Chapter 5

Retrieval Effectiveness

$$\text{Precision} = \frac{\text{Relevant Retrieved}}{\text{Total Retrieved}}$$

$$\text{Recall} = \frac{\text{Relevant Retrieved}}{\text{Total Relevant}}$$

$$\text{F1-score} = \frac{2(P \times R)}{P + R}$$

$$\text{Accuracy} = \frac{\text{Relevant Retrieved} + \text{Irrelevant NOT retrieved}}{\text{Total document Repository}}$$

$$\text{Emeasure} = 1 - \text{F1 score}$$

$$\text{Fallout} = \frac{\text{Retrieved Irrelevant}}{\text{Total Irrelevant}}$$

$$\text{Noise} = 1 - \text{Precision}$$

$$\text{Silence (Miss)} = 1 - \text{Recall}.$$

Chapter 6:
Relevance feedback & Query Expansion
(Query operations)

* Query operations in IR are performed using different approaches to express user information needs more precisely and to improve Retrieval effectiveness. The main approaches include:
  - Boolean
  - Probabilistic, &
  - Vector Space
  - Fuzzy/ Approximate Methods

* Query operation is Action on the Query. (Add, Remove, Reweight terms)

Boolean Approach

- Uses logical operators (AND, OR, NOT) to combine query terms.
- Document either Match or don't Match the query (Binary Relevance)
- pros: Simple & precise for complex conditions
- Cons: No Ranking: Results are either retrieved or not, which may limit usefulness.

Example: Education and Technology NOT online

Vector Space Approach:

- Represents queries and documents as vectors in a multidimensional space.
- Terms are assigned weights based on importance (eg. TF-IDF)
- Documents are ranked based on similarity measures. Usually Cosine Similarity.
- pros: provides ranked results and partial matching
- Cons: Computationally more complex than Boolean.

Example: Query Vector: (0.2, 0.5, 0.3), document Vector (0.1, 0.4, 05)

* Relevance feedback (User, pseudo, interactive): Both a query operation (expanding/Reforming the query) and an approach (Manual, automatic, iterative) to improve Retrieval.

⇒ Relevance feedback in IR is a special type of Query operation where the system improves search results using information about which documents are Relevant or not. It is a technique for Query Reformulation, expansion and an approach to improving Retrieval effectiveness.

1. User Relevance feedback / Explicit feedback)

Defn: The user explicitly Marks Retrieved documents as Relevant or non-Relevant.

Query operation involved:- Adds or reweights terms from Relevant documents → Query expansion / reformulation.

Approach: iterative Refinement guided by actual User judgements.

Pros: Highly Accurate because it reflects real user intent.

Cons: Requires User effort and time.

Example: After Retrieving 20 articles, The user Marks 5 as relevant; the System uses terms from these 5 to refine query.

Chapter 6. Relevance...

2. Pseudo-Relevance feedback :- also called Blind feedback.

defn. The system assumes the top-ranked documents from the initial query are relevant without asking the user.

Query operation involved: Automatically selects important terms from top documents → query expansion

Approach : System-driven, automatic refinement

Pros: No user effort required; can improve Recall

Cons: If top documents are irrelevant, precision may decrease

Example :- Top 10 documents from a search on "Climate Change" are used to expand the query with terms like "Carbon emissions" "global warming" etc.

3. User's (interactive / iterative) Relevance feedback.

defn: Combine user input and iterative query refinement

Query operation involved: Query is reformulated based on user judgements multiple times → iterative Expansion / term reweighting

Approach: interactive Method that repeatedly improves the query.

Pros: Highly effective; aligns with actual user intent

Cons: Requires multiple interactions and time.

Example: A user repeatedly marks relevant documents and the system adjusts the query after each iteration until results are satisfactory.

User Relevance feedback (URF)

Meaning: This is the traditional/classical relevance feedback process. The system retrieves a set of documents → the user marks which documents are relevant or not.
- The system updates the query.

Key Characteristics

Aspects	Description
Who initiates?	- The system first presents results; the user provides feedback afterwards.
User Role	- User labels documents (Relevant / NOT Relevant)
System Role	- System adjusts the query using Rocchio / similar algorithms.
Interaction level	- low to moderate → user mainly gives labels.
Focus	- Improve precision and Recall by modifying the query.

Example:
- User searches: "Water pollution Ethiopia"
- System returns a list
- User marks: Relevant ✓ 5 documents
  NOT Relevant ✗ 3 documents
- System refines the query:
  "Water Contamination, river pollution, Environmental Management Ethiopia"

Chapter 6: Relevance...

Interactive Relevance feedback (IRF)

Meaning: A more dynamic, iterative and user-system cooperative process. The system and the User work together through multiple rounds, not just one.

Key Characterstics

Aspect	Description
✓ Who initiates?	✓ Both User and systems. User may directly Modify query terms.
✓ User Role	- User interacts with the interface! Select terms, expands query, browses clusters, adjust sliders. etc
✓ System's Role	- provides Suggestions: term clouds, Categories, filters, previews.
✓ Interaction Level	- High - user controls the search process step by step
- Focus	- Helping the User explore, refine, and Understand the Search Space.

Core Difference

features	User	Interactive
✓ interaction styles	✓ one-way feedback	✓ Two-way, Continous interaction
✓ User engagement	✓ Limited	✓ High
✓ Query changes	✓ Mostly Automatic	✓ User + system together
✓ Rounds	✓ Usually 1-2	✓ Many rounds
✓ Goal	✓ improves retrieval accuracy	✓ improve the search experience, exploration & contro
✓ Tools used	✓ only document labeling	✓ Suggestion tools filters, Sliders Visualizations

In Simple terms:-

User Relevance feedback

"User marks documents as Relevant - system adjusts the query.

Mostly labeling

Interactive Relevance feedback:

"User and system work together step by step to refine the search."

Exploration + Collaboration + multiple interactions

⇒ User Relevance feedback

User: These 4 docs are Relevant

System: OK, I will modify the query

⇒ interactive Relevance feedback

User: Add term: "River Contamination"

Remove term: "pollution Control strategies"

Filter by Year

Select Cluster "Ethiopia Water Studies"

System Continously updates

Chapter 6 Relevance ...

Techniques for enhancing Retrieval effectiveness in IR.

* Improving Retrieval effectiveness means improving how well an IR system retrieves relevant documents and filters out non-relevant ones.
- The goal is to Maximize precision, Recall and overall User satisfaction

Below are the major techniques used to improve Retrieval effectiveness

1. Query Reformulation is a core technique for enhancing Retrieval effectiveness.

Query Reformulation is the process of modifying, expanding or refining a user's original query to improve the quality of Retrieval results.
- It helps ensure that the system retrieves documents that more accurately match the users true information need.

Why? Users often submit:
- Short queries
- ambiguous terms
- Incomplete descriptions
- poorly expressed needs

Therefore, Retrieval systems use reformulation to:-
- Clarify the query
- Add useful terms
- Remove noisy terms
- Improve precision & Recall

Techniques of Query Reformulation
= Techniques for enhancing Retrieval effectiveness.

a) Manual Query Reformulation (User driven)

* The user modifies the query manually :-

Example include :-

- Using boolean operators [AND, OR & NOT]
- Adding Synonyms
- Removing irrelevant key words
- Narrowing & broadening Search terms

- Good when the user knows the domain

b) Automatic Query Reformulation (System driven)

* The system changes the query w/o user intervention

Common Methods :-

- Query expansion (QE)

Adding Synonyms, related terms, stems
Using thesauri (WordNet), ontologies, term co-ocurrence

- Query Reduction

Removing unhelpful or noisy terms
Keeping only the most discriminative terms.

- Spelling correction / Normalization

Correcting Spelling errors.
Handling variants (Color → colour)

- Feedback-based Reformulation

User Relevance feedback (URF) Explicit
pseudo-Relevance feedback (Implicit)
Interactive feedback (IRF)

Chapter 6 Relevance...

Example of Query Reformulation

1. User Query "Virus"
   System expands to :-
   → Virus + Infection, Malware, antiviral, pathogen.

2. User Query: "Education technology Ethiopia"
   Reformulated to:
   Education technology + ICT in Education, digital learning, Ethiopia, e-learning

---

* Query formulation Vs Query Reformulation are two different stages in the search process in IR.

+ Query Formulation is the process by which a user creates the first/initial query to express their information need in a search system.

Characteristics

- Done before any search results appear
- Expresses the user's understanding of the topic
- May be incomplete, Vague or ambiguous
- Often short (Average 2-3 words)

Activities Involved.

- Selecting keywords
- Choosing filters
- Using boolean operators
- Typing a natural-language question

* A Thesaurus in IR is a structured, controlled Vocabulary that lists Concepts and shows the relationship b/n them. (Synonyms, broader terms, Narrower terms, and Related terms)

- It helps both Users and IR Systems overcome problems of Vocabulary Mismatch, ambiguity, and inconsistent terminology.

A thesaurus in IR is a tool that provides :-

- ✓ Synonyms (eg. "Car" - "automobile"
- ✓ Broader Term (BT)
- ✓ Narrower Term (NT)
- ✓ Related term (RT)
- ✓ prefered descriptors or standard terms

* In IR, It Serves as a bridge b/n the user language and the indexing language.

Why Thesauri improves Retrieval Effectiveness.

- ✓ Handling Synonymy
- ✓ Handling polysemy (multiple meaning)
- ✓ Standardized Vocabulary
- ✓ Supporting Query expansion
- ✓ Guiding novice users