Unit 3: Evaluation and Visualization (Q1 & Q2)

Rank	Question / Topic	Occurrences	Max Marks	Sources
1	**Performance Evaluation Measures (Precision & Recall):** Define, Explain trade-offs, or Calculate using a given dataset.	**6**	9 1	, 2 , 3 , 4 , 5 , 6
2	**User-Oriented Measures:** Explain user-oriented measures in performance evaluation.	**5**	9 1	, 2 , 3 , 7 , 5 , 6
3	**Visualization Techniques:** Query specification/techniques in information visualization & Interface support.	**5**	9 1	, 8 , 3 , 4 , 5 , 6
4	**Advanced Metrics (MRR, NDCG, F-Score):** Explain these specific terms with examples.	**4**	9 1	, 2 , 9 , 5 , 6
5	**Relevance Judgment:** Explain Relevance Judgment, Group Relevance, and Pseudo Relevance Feedback.	**2**	9 7	, 10

Exam Cheat Sheet: Which one to write?

If the Question asks for...	You write...
"Basic measures" or **"Precision/Recall"**	Precision, Recall, and the Trade-off graph.
"User-Oriented Measures"	Coverage, Novelty, Relative Recall, Recall Effort.
"Alternative Measures"	MRR, NDCG, F-Score, E-Measure.
"Single Value Summaries"	F-Score, E-Measure (sometimes Average Precision).
"Evaluation for Ranked Retrieval"	MRR, NDCG.

Q1) Define Precision and Recall & F1 SCORE. Explain the trade-off between them.

- **Precision:**
  - Precision is the ratio of the number of **relevant documents retrieved** to the **total number of documents retrieved**.
    - It measures the accuracy of the search process.

      - **Formula:**

$$Precision = \frac{|Relevant \cap Retrieved|}{|Retrieved|} = \frac{A}{A+B}$$

'Where A = Relevant retrieved, B = Irrelevant retrieved) 1 .

- **Recall:**
  - Recall is the ratio of the number of **relevant documents retrieved** to the **total number of relevant documents in the database** (collection).

    - It measures the completeness of the search.

      - **Formula:**

$$Recall = \frac{|Relevant \cap Retrieved|}{|TotalRelevant|} = \frac{A}{A+C}$$

'Where A = Relevant retrieved, C = Relevant not retrieved) 1 .

• **F1 Score (F-Measure):**
The F-Score is the **harmonic mean** of Recall and Precision.
It provides a single score that balances both precision and recall. A high F-measure ensures that both precision and recall are reasonably high.

◦ **Formula:**

$$F = \frac{2 \cdot P \cdot R}{P + R}$$

Characteristics: F=1 indicates all ranked documents are relevant; F=0 indicates no relevant documents were retrieved.

2. Trade-off Between Precision and Recall
There is generally an **inverse relationship** (trade-off) between Precision and Recall.

• **The Conflict:** As Recall increases, Precision tends to decrease, and as Precision increases, Recall tends to decrease.

• **Explanation:**

◦ To achieve **High Recall** (find *every* relevant document), the system must retrieve a large volume of documents. By broadening the search scope to catch hard-to-find relevant items, the system inevitably retrieves more irrelevant "junk" documents, thereby lowering **Precision**.

◦ To achieve **High Precision** (ensure *only* relevant documents are returned), the system must be very selective. By filtering out anything doubtful, the system will likely miss some relevant documents, thereby lowering **Recall**.

• **Graphical Representation:** In standard evaluation graphs, precision usually drops to 0% for recall values above 50% because further increases in recall require retrieving significantly more non-relevant documents. Ideally, a system wants to maximize the F-measure to find the best possible compromise between the two.

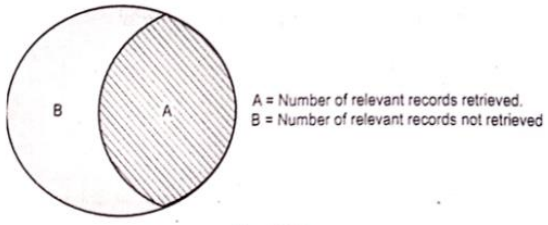

Fig. Q.3.2

$$\text{Recall} = \frac{A}{A+B} \times 100\ \%$$

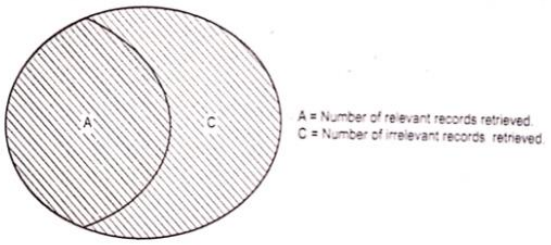

Fig. Q.3.3

$$\text{Precison} = \frac{A}{A+C} \times 100\ \%$$

---

q2) Why are the performance evaluation measures needed in IR system

1. Validating Design Decisions

An IR system involves making several complex architectural decisions before it is operational. Evaluation is necessary to verify if these decisions result in an effective system. These decisions include:
- Finding the internal representation of documents.
- Deciding the file structure to save documents inside the system.
- Selecting the mathematical model for matching documents with the query (e.g., Vector Space vs. Boolean).

2. Functional and Error Analysis

Evaluation is required to perform two specific types of analysis during the testing phase:
- **Functional Analysis:** Assessing whether the system meets the core requirements of retrieving information.
- **Error Analysis:** Identifying specific failures in the system to understand *why* irrelevant documents are

retrieved or relevant ones are missed.

3. Measuring Relevance (The "Heart" of IR)

Measuring the relevance of documents with respect to a user's query is considered the "heart" of Information Retrieval. Since "relevance" is subjective, formal measures are needed to quantify how well the system's "actual retrieval set" matches the "set of relevant items" in the database,.

4. Efficiency Assessmend

Beyond just finding the right documents (effectiveness), evaluation is needed to measure the system's efficiency in terms of resources:

- **Response Time:** Measuring the time between query submission and the answer.
- **Space:** Measuring the memory or storage required for indexing structures.

5. Balancing the Trade-off

Evaluation measures are needed to visualize and manage the inherent trade-off between **Precision** (accuracy) and **Recall** (completeness). Without measuring these, a developer cannot know if the system is too strict (missing information) or too loose (providing too much junk),.

Q3) Explain user-oriented measures in performance evaluation.

These measures differ from standard system-centered measures (like Precision and Recall) because they focus on the **user's prior knowledge** and the **effort** required by the user to find information, rather than just the statistical accuracy of the system.

User-Oriented Performance Measures

To understand these measures, we assume the user already knows about some relevant documents (Rk) and is looking for new ones (Ru). The retrieval system returns an answer set (A).

1. Coverage Ratio

• **Definition:** This measures how well the system finds the relevant documents that the user **already knows** exist. It indicates the system's ability to cover known ground.

• **Explanation:** If a user knows about 10 specific papers on a topic and the search engine only returns 2 of them, the coverage is low.

- **Formula:**

$$Coverage = \frac{\text{Known Relevant Documents Retrieved}}{\text{Total Known Relevant Documents}}$$

$$Coverage = \frac{|R_k \cap A|}{|R_k|}$$

(Where R_k is the set of relevant documents known to the user, and A is the retrieved set) [1].

2. Novelty Ratio

• **Definition:** This measures the system's ability to find **new** information (documents previously unknown to the user)

• **Explanation:** A user usually searches to find things they *don't* know. If the system retrieves 10 relevant documents, but the user had already read 9 of them, the search has low novelty value. A high novelty ratio means the system is effective at discovering new information.

• **Formula:**

$$Novelty = \frac{\text{Unknown Relevant Documents Retrieved}}{\text{Total Relevant Documents Retrieved}}$$

$$Novelty = \frac{|R_u \cap A|}{|R \cap A|}$$

(Where R_u is the set of relevant documents unknown to the user, and $R \cap A$ is the total relevant documents retrieved) [1] [2].

3. Relative Recall

• **Definition:** This is used when the absolute total number of relevant documents in the database is unknown (which is true for the Web). It compares the success of the search against the user's **expectations**.

• **Explanation:** Instead of calculating recall based on the entire database (which is impossible to count), it calculates it based on the number of relevant documents the user *expected* to find.

• **Formula:**

$$RelativeRecall = \frac{\text{Relevant Documents Retrieved}}{\text{Expected Number of Relevant Documents}}$$

(Note: This is a subjective measure based on user feedback) [2].

4. Recall Effort

• **Definition:** This measures the amount of work the user has to do to find the relevant information.

• **Explanation:** It is the ratio of relevant documents found to the total number of documents the user had to examine (read/scan) in the result list.

• **Formula:**

$$RecallEffort = \frac{\text{Relevant Documents Retrieved}}{\text{Total Documents Examined}}$$

(Note: This is mathematically similar to Precision, but conceptually focuses on the "effort" or "noise" the user deals with) [2].

Summary of Terms (for Set Notation)

• Rk **(Known Relevant):** Documents the user already knows are relevant.

• Ru **(Unknown Relevant):** Relevant documents the user has not seen before.

• A **(Answer Set):** The documents retrieved by the system.

• U**:** The universe of all documents

---

q4) **"Alternative measures"** or **"Single Value Summaries.":** Performance Evaluation

1. F-Score (Harmonic Mean)

Concept: Since maximizing both Precision and Recall simultaneously is difficult (the trade-off), the F-Score provides a single score that balances both.

• **Formula:**

$$F = \frac{2 \times P \times R}{P + R}$$

(Where P = Precision, R = Recall)

Key Point: It is high only when **both** Precision and Recall are high. If either is zero, the F-Score is zero.

2. Mean Reciprocal Rank (MRR)

Concept: Used primarily for systems that return a ranked list of answers (like Google). It evaluates how far down the list the **first** relevant document appears.

• **Formula:**

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

(Where $|Q|$ is the number of queries and $rank_i$ is the rank position of the first relevant document for the i-th query).

Example: If the first relevant document is at rank 1, score is 1. If at rank 2, score is 0.5. If at rank 10, score is 0.1. Higher is better.

3. NDCG (Normalized Discounted Cumulative Gain)

Concept: Unlike Precision/Recall which are binary (Relevant vs. Not Relevant), NDCG handles **Graded Relevance** (e.g., Highly Relevant, Partially Relevant, Irrelevant). It also accounts for the **position** of the result—highly relevant documents appearing earlier get a higher score.

Components:

CG (Cumulative Gain): Sum of relevance scores.

DCG (Discounted CG): Penalizes highly relevant documents if they appear lower in the list (divides score by log of rank).

NDCG: Normalizes the score so it is between 0 and 1,

4. E-Measure

Concept: A variant of the F-measure that allows the user to specify whether they care more about Precision or Recall using a parameter β.

Behavior:

• **Formula:**

$$E = 1 - \frac{1+\beta^2}{\frac{\beta^2}{R}+\frac{1}{P}}$$

• **Behavior:**

◦ $\beta > 1$: Emphasizes **Recall**.

◦ $\beta < 1$: Emphasizes **Precision**.

◦ $\beta = 1$: Equal weight (equivalent to complement of F-measure) [1], [4].

---

q5) **Single Value Summaries (Trade-off Measures)**

These measures are used because comparing systems based on two conflicting metrics (Precision and Recall) is difficult. A "Single Value Summary" combines them into one number to make it easier to rank the performance of different algorithms.

1. F-Measure (F-Score)

This is the most frequently asked single-value metric in your exams. It resolves the trade-off by taking the **harmonic mean** of Precision and Recall.

• **Definition:** It is a measure that combines precision and recall. It is high only when **both** precision and recall are high.

• **Formula:**

$$F(j) = \frac{2 \cdot P(j) \cdot r(j)}{P(j) + r(j)}$$

(Where $P(j)$ is Precision and $r(j)$ is Recall at the j-th document) [2].

• **Why Harmonic Mean?** The harmonic mean is used instead of the arithmetic mean (average) because the arithmetic mean is affected by large values. If a system has 100% Recall but 1% Precision, the arithmetic average would still look decent (50%), but the Harmonic mean will properly penalize the system for the low precision.

• **Characteristics:**

- **F = 0:** No relevant documents were retrieved.
- **F = 1:** All ranked documents are relevant.
- It represents the best possible compromise between recall and precision.

2. E-Measure

This is a variant of the F-measure that allows the user to specify whether they care more about Precision or Recall using a parameter 'b'.

• **Definition:** A measure that combines recall and precision while allowing the user to specify the relative importance of one over the other.

- **Formula:**

$$E(j) = 1 - \frac{1 + b^2}{\frac{b^2}{r(j)} + \frac{1}{P(j)}}$$

- **The Parameter 'b' (User Preference):**
  - $b = 1$: Acts as the complement of the F-measure (Equal weight to Precision and Recall).
  - $b > 1$: The user is more interested in **Recall** (finding everything, even if it means some junk).
  - $b < 1$: The user is more interested in **Precision** (accuracy, minimizing junk), 3 .

---

q6) Define Query. What are various techniques used to specify query in information visualization?

A **Query** is a formal expression of the user's information need. It serves as the bridge between the user's vague requirement for information and the system's database. In Information Retrieval (IR), a query is typically a set of keywords, boolean operators, or natural language sentences that the system processes to retrieve relevant documents.

• **Key Characteristic:** In visualization systems, queries allow users to limit the view of data to only the items that satisfy specific criteria, thereby filtering out "noise" and focusing on relevant information

Various Techniques Used to Specify Query (Interface Support)

According to the syllabus and Shneiderman's classification for Human-Computer Interaction (HCI) in IR systems, there are five primary techniques (or styles) used by users to specify queries in an information visualization interface:

1. Command Line Interface

• **Description:** This provides a means of expressing instructions to the computer directly using function keys, single characters, abbreviations, or whole-word commands.

• **Usage:** It is often the only way to communicate with older systems (e.g., via Telnet). It requires the user to know specific syntax (like Boolean queries using `AND`, `OR`, `NOT`).

• **Pros/Cons:** It is powerful for expert users but counter-intuitive and difficult for beginners.

2. Menu Selection

• **Description:** The set of available query options is displayed on the screen, and the user selects them using a mouse, numeric keys, or alphabetic keys.

• **Usage:** These visible options rely on **recognition** rather than **recall** (users don't need to memorize commands). Menus can be nested hierarchically or grouped logically to guide the user.

3. Form-Fills and Spreadsheets

• **Description:** The display resembles a paper form with slots (fields) to fill in. It is primarily used for data entry but is very effective for data retrieval (Advanced Search).

• **Usage:** The user enters text into specific fields (e.g., "Author Name", "Date Range"). It helps users by providing a structured context for the query.

4. Natural Language

• **Description:** The user types a query in normal human language (e.g., "Find all documents about cancer treatments").

• **Usage:** While ideal for users, it is very difficult for a machine to understand due to ambiguity and semantic complexity. The system must parse the sentence to extract meaning.

5. Question/Answer (Query Dialogue)

• **Description:** This is a simple mechanism where the system guides the user through a series of questions to narrow down

the search.
• **Usage:** The user is asked a question, provides an answer, and the system asks the next relevant question. It is easy to learn but can be limited in functionality and slower to execute.

Additional Query Specification Concepts

• **Boolean Queries:** Users join query terms using logic operators (`AND`, `OR`, `NOT`). While precise, users often struggle with the logic (e.g., using `AND` when they mean `OR`).

• **Faceted Queries:** This technique solves the complexity of Boolean queries by presenting specific categories (facets) like "Price", "Brand", or "Color" that users can click to refine results, common in e-commerce.

Q7) Define and explain Interface support for search process related to visualization in information system

Definition: Interface support for the search process refers to the visual and functional design decisions made to structure the user's interaction with an Information Retrieval (IR) system. It determines how queries are entered, how results are displaye d, and how users navigate through large volumes of information without becoming overwhelmed,

The goal is to arrange various kinds of information on the computer screen to support the user's cognitive processing during a search.

1. Key Design Variables

the interface designer must address four critical variables to determine the appropriate visualization support,:

1. **User Expertise:**
   ◦ Are users comfortable with **Boolean operators** (e.g., AND, OR) or do they prefer **Natural Language**?
   ◦ Do they need a simple search box or a high-powered interface with many filters?
2. **User's Information Need:**
   ◦ Does the user want just a "taste" (brief summary) or "comprehensive research"?
   ◦ Should the results be displayed as short abstracts or extensive details?
3. **Type of Information:**
   ◦ Is the data made up of structured fields (like a database), full text, or navigation pages?
4. **Information Volume:**
   ◦ How much information is being searched? The interface must prevent the user from being overwhelmed by the number of retrieved documents.

2. Interface Support for Search Process

- Search process gives results of documents which are relevant to the user query.
- These results need to be presented properly to the user.
- Results can be sorted by order of relevance.
- Another parameter can be frequently used documents.
- Interface should allow tracing of user relevance parameters.
- Interface should allow tracking of implicit as well as explicit user relevance parameters.

2. Visualization & Layout Techniques

To support the search process effectively, systems use specific layout strategies to manage windows and display content contextually.

A. Window Management

The designer must choose how to arrange information on the screen:

• **Tiled Windows:** Windows are laid out in predefined positions (e.g., side-by-side) and never overlap. This allows users to see the search list and the document details simultaneously.

• **Overlapping Windows:** Windows can cover each other, simulating a messy desk. This saves screen space but might hide important context.

B. Specific Visualization Layouts

The sources highlight two specific frameworks used to support the search process visually,:

1. **The InfoGrid Layout:**
   ◦ **Concept:** A "monolithic" layout designed for information access.

- **Structure:** The screen is divided into specific areas.
  - **Left Hand Side:** Devoted to retrieval and control (e.g., search parameters, thumbnails of documents, search paths/history).
  - **Right Hand Side:** Devoted to viewing the actual document text or results.

• Fig. Q.15.1 shows InfoGrid layout.

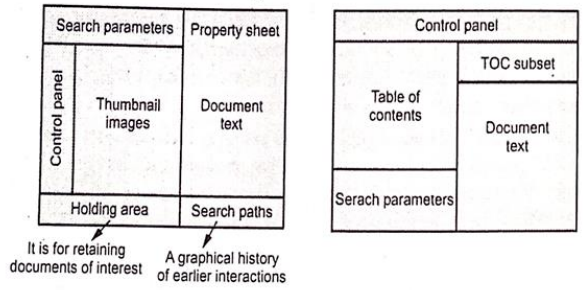

Fig. Q.15.1 InfoGrid layout

- **Benefit:** It keeps the user's tools (controls) and the task (reading) visible at the same time.

2. **The SuperBook Layout:**
   - **Concept:** This uses the structure of a large document (like a book) to provide context.
   - **Structure:**
     - **Left Pane (TOC):** Displays the **Table of Contents**. It uses visual indicators (hits) to show how many times the search term appears in each chapter or section.
     - **Right Pane (Text):** Displays the selected document text.
   - **Benefit:** Users can manipulate the TOC to expand/contract sections. This allows users to see "term hits" in context — knowing *where* in the book the information is (e.g., "The word 'virus' appears 50 times in Chapter 3 but only once in Chapter 1")

---

q8) Define and explain following terms- [9] i) MRR ii) NDCG iii) E-measure

i) MRR (Mean Reciprocal Rank)

Definition: MRR is a statistical measure used to evaluate systems that produce a ranked list of possible responses to a query. It focuses on the rank of the **first** relevant document found.

Concept:

In many searches (like looking for a specific fact), the user only cares about the *first* correct answer. If the first correct result is at rank 1, the system is perfect. If it is at rank 10, the system is poor.

- The "Reciprocal Rank" for a single query is $\frac{1}{\text{Rank of first relevant document}}$.

MRR is the average (mean) of these reciprocal ranks across multiple queries.

- **Formula:**

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

  - Where $|Q|$ is the total number of queries.
  - $rank_i$ is the rank position of the *first* relevant document for the i-th query.

- **Example:**
  - Query 1: First relevant doc at Rank 1 → Score = 1/1 = 1.
  - Query 2: First relevant doc at Rank 2 → Score = 1/2 = 0.5.
  - Query 3: First relevant doc at Rank 5 → Score = 1/5 = 0.2.
  - **MRR** = $(1 + 0.5 + 0.2)/3 = 0.56$ 1.

ii) NDCG (Normalized Discounted Cumulative Gain)

Definition: NDCG is a measure of ranking quality that calculates the gain (usefulness) of a document based on its position in the result list. Unlike Precision (which is binary: Relevant/Not Relevant), NDCG supports **graded relevance** (e.g., Highly Relevant, Partially Relevant, Irrelevant).

Key Characteristics:

Graded Relevance: It acknowledges that some documents are "more relevant" than others.

Position Matters: Highly relevant documents appearing lower in the list should be penalized (discounted).

How it is calculated (The Process):

CG (Cumulative Gain): The sum of relevance scores of retrieved documents.

2. **DCG (Discounted Cumulative Gain):** The relevance score is divided by the log of its rank position. This reduces the value of a relevant document if it is found far down the list.

$$DCG = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2(i)}$$

3. **NDCG (Normalized):** To compare queries with different numbers of results, DCG is divided by the "Ideal DCG" (IDCG), which is the DCG of a perfect ordering.

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

- The value ranges from 0 to 1, with 1 being a perfect ranking [1], [2].

iii) E-measure

Definition: The E-measure is a parameterized performance metric that combines Precision (P) and Recall (R). It allows the user to specify their preference (whether they care more about accuracy or completeness) using a parameter b.

Concept: It essentially acts as the complement to the weighted F-measure ($E = 1 - F$). A lower E-value indicates better performance (closer to 0 is good, whereas for F-score, closer to 1 is good).

- **Formula:**

$$E = 1 - \frac{1 + b^2}{\frac{b^2}{R} + \frac{1}{P}}$$

- **The Parameter 'b' (User Preference):** The variable b reflects the relative importance of recall vs. precision:
  - $b = 1$: Equal weight to Precision and Recall (Acts as the complement of the standard F1-measure).
  - $b > 1$: The user is more interested in **Recall** (finding all relevant documents is more important than precision).
  - $b < 1$: The user is more interested in **Precision** (avoiding errors is more important than finding everything) [3], [4], [5].

Q9) What is relevance Judgement? Explain the term group relevance judgements, pseudo relevance feedback.

1. What is Relevance Judgment?

Relevance judgment is the process of assessing whether a retrieved document satisfies the user's information need expressed in their query. It is considered the "heart" of Information Retrieval (IR) evaluation because relevance is the prim ary criterion for success.

- **Definition:** It is the measurement of the correspondence between a document and a query. If a document provides the information the user was looking for, it is judged as "Relevant"; otherwise, it is "Not Relevant",.
- **Subjectivity:** Relevance is inherently subjective. Two different users might judge the same document differently for the same query based on their background knowledge, the time of search, or their specific intent (e.g., one wants a summary, another wants a detailed report).
- **Types of Judgment:**
  - **Binary:** Relevant (1) vs. Non-Relevant (0).
  - **Graded:** Highly Relevant, Partially Relevant, Not Relevant (used in measures like NDCG).

2. Group Relevance Judgments

While individual relevance judgment is subjective, **Group Relevance Judgments** are used to establish a reliable "ground truth" or "gold standard" for evaluating and comparing IR systems (standard in benchmarking conferences like TREC).

- **Concept:** Since a single user's judgment might be biased or inconsistent, a **group of experts** or a panel of assessors is used to judge the relevance of documents for a set of standard queries.
- **Purpose:** To create a standard dataset (Test Collection) where the "correct" answers are agreed upon by a group. This allows developers to calculate Precision and Recall scores that are statistically significant and not just based on one person's opinion.
- **Method:**
  - Multiple assessors review the documents.
  - Consensus is reached (e.g., majority vote).
  - This "Group Judgment" becomes the benchmark against which the IR system's output is measured.

3. Pseudo Relevance Feedback (Blind Feedback)

Pseudo Relevance Feedback (Pseudo RF), also known as **Blind Relevance Feedback**, is an automated technique used to improve search results without requiring any actual input or effort from the user.

- **Definition:** It is a method where the system assumes that the top-ranked documents returned by an initial query are relevant, even though the user has not confirmed them. The system then uses these documents to refine the query automatically.
- **The Process (Algorithm):**
  - **Initial Search:** The system runs the user's original query and retrieves a ranked list of documents.
  - **Assumption:** The system **blindly assumes** that the top k documents (e.g., the top 10) are relevant (hence the term "Pseudo" or "Fake" relevance).
  - **Expansion:** The system analyzes these top k documents to identify prominent terms (keywords) that appear frequently.
  - **Reformulation:** These new terms are added to the original query (Query Expansion), or the weights of original terms are adjusted.
  - **Final Search:** The system executes the modified query to produce a final, improved ranked list for the user.
- **Advantages:**
  - It does not require any user interaction (unlike standard Relevance Feedback where the user must manually tick boxes).
  - It often improves recall by finding documents that use synonyms of the query terms found in the top results.
- **Disadvantages (The Drift Problem):**
  - It works well for "good" initial queries where the top results are actually relevant.
  - If the initial query is "bad" and retrieves non-relevant documents (noise) in the top k, the system will expand the query with junk terms. This leads to **Query Drift**, where the search moves away from the user's actual intent.

1. Example of Relevance Judgment

Scenario: A user is searching for information about **"Java"** on a search engine.

• **The Query:** "Java history".

• **The Result:** The system retrieves a document titled *"The History of Coffee Production in Java, Indonesia."*

• **The Judgment:** The user, who is a computer science student looking for the programming language, reads the title and snippets. They decide this document is **"Not Relevant"** because it does not satisfy their specific information need.

- *Key Point:* This is a subjective decision made by the specific user at that specific moment.

2. Example of Group Relevance Judgment

Scenario: A company is testing a new search algorithm and needs to calculate its **Precision** and **Recall** accurately. They cannot rely on just one person's opinion because that person might be biased or make a mistake.

• **The Process:**

1. They select a standard query: **"Global Warming effects"**.
2. They hire a **panel of 3 to 5 subject matter experts** (assessors).
3. This group reviews Document A. Two experts say "Relevant," one says "Not Relevant."
4. **Consensus:** Through majority vote or discussion, the group decides Document A is officially "Relevant."

• **The Outcome:** This group decision becomes the **"Ground Truth"** or "Gold Standard." The system is then graded on whether it retrieves Document A. If it fails to retrieve it, it is a system error, regardless of what a casual user might think.

3. Example of Pseudo Relevance Feedback (Blind Feedback)

Scenario: A user searches for **"Smartphones"** but the initial results are too broad.

• **Step 1 (Initial Search):** The system runs the query "Smartphones" and finds 1,000 documents.

• **Step 2 (The "Blind" Assumption):** The system blindly assumes that the **top 5 ranked documents** are relevant, without asking the user to confirm.

◦ *Top Doc 1:* "Latest Apple iPhone features..."

◦ *Top Doc 2:* "Samsung Galaxy reviews..."

◦ *Top Doc 3:* "Android vs iOS operating systems..."

• **Step 3 (Feature Extraction):** The system analyzes these top 5 documents and finds frequent terms like **"Apple"**, **"Samsung"**, **"Android"**, and **"4G"**.

• **Step 4 (Query Expansion):** The system automatically adds these terms to the user's query.

◦ *New Internal Query:* "Smartphones Apple Samsung Android 4G"

• **Step 5 (Final Result):** The system runs this new query to get a better, more specific list of results for the user.

◦ *Risk (The Drift Problem):* If the top document was actually an article about *"Smartphones causing lack of sleep,"* the system might add terms like "Sleep" and "Insomnia" to the query, drifting the search away from technology and toward health, which might not be what the user intended,.