Unit 3 (ISR)

1. **Define Precision and Recall. Give example of each and justify its use in evaluating IR system.**

**Precision**

**Definition**: Precision is the proportion of retrieved documents that are relevant. It measures the accuracy of the results returned by the Information Retrieval (IR) system.

**Example**:  
Suppose a search engine retrieves 10 documents for a query, and 7 of them are relevant while 3 are irrelevant.

* **Precision** = 710=0.7\frac{7}{10} = 0.7 or 70%.

**Recall**

**Definition**: Recall is the proportion of relevant documents that are successfully retrieved. It measures the completeness of the system in finding all the relevant documents.

**Example**:  
In the same scenario, suppose there are 12 relevant documents in the entire database, but the system retrieved only 7.

* **Recall** = 712≈0.58\frac{7}{12} \approx 0.58 or 58%.

1. **In information retrieval, if q is the information request and a set of relevant documents for query q is Rq = (d3, d5, d9, d25, d39, d44, d50, d70, d80, d120). Consider new retrieval algorithm has been designed and has been evaluated for information request q returns, ranking of the documents in the answer set is as follows.**
2. **d120**
3. **d84**
4. **d50**
5. **d6**
6. **d8**
7. **d9**
8. **d58**
9. **d129**
10. **d143**
11. **d25**
12. **d38**
13. **d48**
14. **d230**
15. **d113**
16. **d3**

**Ans :**

Given the information request **q** and the set of relevant documents for **q** as:

Rq={d3,d5,d9,d25,d39,d44,d50,d70,d80,d120}

The retrieval algorithm ranks documents as follows:

1. d120
2. d84
3. d50
4. d6
5. d8
6. d9
7. d58
8. d129
9. d143
10. d25
11. d38
12. d48
13. d230
14. d113
15. d3

**Precision and Recall Calculation:**

To calculate **Precision** and **Recall**, we'll first check which of the retrieved documents are relevant (from **Rq**).

**Relevant Documents in the Retrieval:**

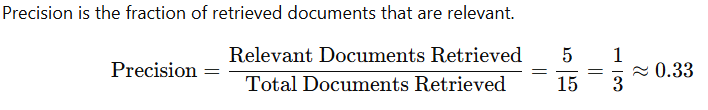
From the given ranking, the relevant documents (from **Rq**) are:

* **d120** (Rank 1)
* **d50** (Rank 3)
* **d9** (Rank 6)
* **d25** (Rank 10)
* **d3** (Rank 15)

Thus, **5 relevant documents** are retrieved.

**Precision Calculation:**

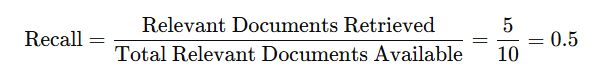
Precision is the fraction of retrieved documents that are relevant.



So, **Precision** = 0.33 or **33%**.

**Recall Calculation:**

Recall is the fraction of relevant documents retrieved out of all relevant documents in the database.



So, **Recall** = 0.5 or **50%**.

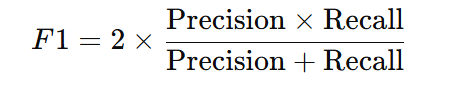
**Summary:**

* **Precision** = 33%
* **Recall** = 50%

1. **What are alternative measures used to evaluate system performance?**

**1) Harmonic Mean (also referred to as F-Score or F1-Score)**

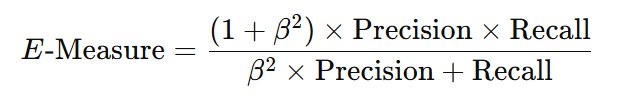
* **Definition**: The Harmonic Mean is used to combine **Precision** and **Recall** into a single metric that balances both. It's particularly useful when there is a need to evaluate how well the system performs with respect to both precision and recall.
* **Formula**:



* **Importance**: The F-Score gives equal importance to both precision and recall, making it ideal when you need a balance between retrieving relevant documents and avoiding irrelevant ones.

**2) E-Measure**

* **Definition**: The **E-Measure** (or **Expected Utility Measure**) is used to combine multiple evaluation metrics into one, taking into account both precision and recall while also incorporating user preferences regarding false positives and false negatives.
* **Formula**:



where β\beta is a parameter that controls the relative importance of precision versus recall. When β=1\beta = 1, it becomes the F-Score.

* **Importance**: The E-Measure can be adjusted based on how much weight you want to assign to precision or recall, making it more flexible for different user needs.

**3) User-Oriented Measures**

* **Definition**: These measures focus on how well the IR system meets the needs and expectations of the user. They often go beyond basic retrieval performance (such as precision and recall) to incorporate aspects such as **user satisfaction** and **the practical utility of the results**.
* **Types of User-Oriented Measures**:
  + **User Satisfaction**: Measures how satisfied users are with the results, which can be influenced by relevance, ranking, and the overall usefulness of the retrieved documents.
  + **Task Performance**: Assesses how effectively the system helps the user complete a task, such as answering a question or finding a specific piece of information.
  + **Click-Through Rate (CTR)**: Measures how often users click on retrieved documents, giving insight into their perceived relevance.
  + **Average Time to Find Relevant Document**: Measures how quickly users can find relevant documents from the search results.
* **Importance**: These measures are particularly useful when focusing on the end-user experience, which is the ultimate goal of any IR system. They provide a more holistic evaluation of system effectiveness based on real user interaction.

**Summary of Alternative Measures:**

* **Harmonic Mean (F-Score)**: Balances **Precision** and **Recall** into one metric.
* **E-Measure**: Combines precision and recall with an additional parameter (β\beta) to adjust their relative importance, useful for different user preferences.
* **User-Oriented Measures**: Focuses on user satisfaction, task completion, and interaction with the system, offering a practical measure of how well the IR system serves real users.

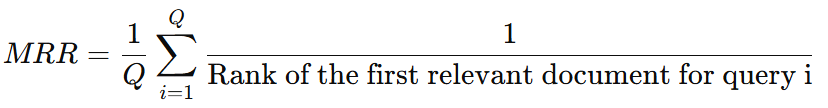
1. **Define and explain following terms (in E Measure) :** 
   1. **MRR**
   2. **NDCG**
   3. **F-score**

**1. MRR (Mean Reciprocal Rank)**

**Definition:**

**Mean Reciprocal Rank (MRR)** is a measure used to evaluate the effectiveness of a ranking system based on how quickly the system retrieves the first relevant document.

**Formula:**



where:

* Q is the total number of queries,
* Rank of the first relevant document\text{Rank of the first relevant document} is the position at which the first relevant document appears in the search results for a particular query.

**Explanation:**

* MRR evaluates how fast the system can return the first relevant document in the ranking.
* It’s especially useful for scenarios where the user is interested in the top result, like in **question-answering systems** or **web search engines**.
* **Higher MRR** values indicate better performance as the first relevant document appears earlier in the ranking.

**2. NDCG (Normalized Discounted Cumulative Gain)**

**Definition:**

**NDCG (Normalized Discounted Cumulative Gain)** is a metric that evaluates the ranking quality of a search engine or IR system. It considers both the **relevance** of documents and their **rank** in the result set. NDCG gives higher weight to relevant documents appearing earlier in the ranking.

**Explanation:**

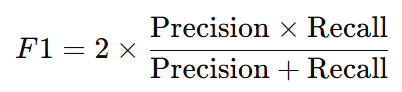
* **NDCG** accounts for both the relevance and position of documents in the ranked list. If a highly relevant document appears near the top, it gets a higher score.
* **Normalization** ensures that the score is bounded between 0 and 1, making comparisons easier across different queries.
* **Higher NDCG** values indicate better ranking performance, where relevant documents are ranked higher.

**3. F-Score (or F1-Score)**

**Definition:**

**F-Score** (or **F1-Score**) is the harmonic mean of **Precision** and **Recall**, balancing the two metrics. It is used when both precision and recall are important, and you want a single measure to evaluate the system’s performance.

**Formula:**



**Explanation:**

* **Precision** is the fraction of relevant documents retrieved out of all retrieved documents.
* **Recall** is the fraction of relevant documents retrieved out of all relevant documents available.
* The **F1-Score** provides a balanced evaluation of both precision and recall. A **higher F1-Score** indicates a better balance between retrieving relevant documents (precision) and retrieving as many relevant documents as possible (recall).

**Summary:**

1. **MRR** evaluates the rank at which the first relevant document appears, with a higher MRR indicating faster retrieval of relevant documents.
2. **NDCG** evaluates the quality of ranked results by considering both relevance and position, with higher NDCG values indicating better-ranked results.
3. **F-Score** combines precision and recall into a single metric, providing a balanced evaluation of the system's ability to retrieve relevant documents while minimizing irrelevant ones.

These measures are commonly used in Information Retrieval to assess how well systems perform, especially in search engines, recommendation systems, and ranking tasks.

1. ***Cross-Fold Evaluation (Cross-Validation) in Information Retrieval***

**Cross-Fold Evaluation** is a technique used to assess the performance of an Information Retrieval (IR) system or machine learning model by dividing the dataset into multiple "folds" and using different combinations of training and testing sets. This helps in evaluating the system's effectiveness without overfitting or biasing the results.

**How It Works:**

1. **Dataset Division**: The dataset is randomly divided into **k** equal parts, or "folds."

Example: If you have a dataset with 100 documents, and you choose **k = 5**, the data will be split into 5 parts, each containing 20 documents.

1. **Training and Testing**:
   * For each fold, one part is used for testing (validation set), while the remaining **k-1** parts are used for training the system.
   * The process is repeated **k** times, each time with a different fold used as the test set.

Example:

* + In the first round, fold 1 is used for testing, and folds 2-5 are used for training.
  + In the second round, fold 2 is used for testing, and folds 1, 3-5 are used for training.
  + This continues for all folds.

1. **Performance Evaluation**: After completing the **k** iterations, the results are averaged to get a more reliable estimate of the system’s performance.

**Example of 5-Fold Cross-Validation:**

* **Dataset**: 100 documents
* **k = 5** (5 folds)
* **Iteration 1**: Fold 1 = test, Folds 2-5 = train
* **Iteration 2**: Fold 2 = test, Folds 1, 3-5 = train
* And so on...

After 5 iterations, the system's performance (precision, recall, etc.) is averaged to get the final result.

**Advantages of Cross-Fold Evaluation:**

1. **Better Generalization**: It helps ensure the system is tested on different portions of the data, reducing the chance of overfitting to a single training set.
2. **Reliable Results**: Since the system is tested multiple times, the evaluation is more robust and accurate.

**Summary**

Cross-fold evaluation (or cross-validation) is a technique where the dataset is split into **k** parts, and the system is trained and tested on different folds. It helps in achieving a more reliable estimate of system performance by averaging results over multiple runs.

1. **Visualization in Information Retrieval Systems**

**Visualization in Information Retrieval Systems** helps to present data in a visual format, making it easier to understand and analyze.

1. **Starting Points**: Visualization identifies patterns in data and aids decision-making.
2. **Document Context**: It shows document details using **term frequency** or **binary representation** (presence/absence of terms).
3. **User Relevance Judgment**: Measures how relevant the results are to the user by factors like:
   * Time spent on a document.
   * Clicking on links.
   * User feedback.
4. **Interface Support for Search**: Presents search results clearly, sorted by relevance or other criteria. It tracks both **explicit** (direct feedback) and **implicit** (behavioral data like clicks) relevance.

**Summary: Visualization helps users interpret search results, improving decision-making and system performance by making data easier to understand.**