Information Retrieval (IR) systems are built to navigate large datasets using relevant input queries. They utilize statistical algorithms to match these queries with reference documents, retrieving the most relevant information based on ranking metrics. You'll see IR in action in popular search engines like Google and even within local organizations for retrieving company documents.

Before rolling out an IR system, it's crucial to evaluate it using specific metrics. These metrics assess how well the system returns relevant search results based on the input queries. Here's a breakdown of some key information retrieval metrics and how they work:

Key Information Retrieval Metrics

1. Precision@k:

- Precision measures how many of the retrieved results are actually relevant. The '@k' refers to the top-k results being evaluated. For example, Precision@5 looks at how many of the first five results are relevant to the search query.

2. Recall@k:

- Recall measures how many relevant items were returned out of all the relevant items available in the dataset. The value of k plays a significant role here; if k equals the total number of relevant documents, recall will be 1. The choice of k should match the needs of the application.

3. F1-Score@k:

- The F1 score is the harmonic mean of precision and recall, offering a balanced view between the two. It's especially useful when you need to consider both metrics.

These metrics are order-unaware, meaning they don't take into account the order of the returned results.

Ranked Metrics

Ranked metrics do consider the order of results and are essential for evaluating IR systems:

1. MAP (Mean Average Precision):

- MAP consists of two parts. First, it calculates the average precision for different values of k (from 1 to N) for a single query. Then, it averages these values across all queries.

2. NDCG (Normalized Discounted Cumulative Gain):

- NDCG looks at two ranks for each item in the database: one assigned by the user, reflecting its relevance, and another provided by the IR system. NDCG compares these rankings to evaluate performance.

By understanding and implementing these metrics, organizations can ensure their IR systems deliver accurate and relevant results.

Let us discuss recall and precision in further view:

Recall and precision, in the context of information retrieval, directly analyze the relevance of the returned results. They have similar workings but different scopes. Precision judges the retrieval system based only on the items returned in the result set. Its formula is

$$Precision@k = \frac{TruePositives@k}{TruePositives@k + FalsePositives@k}$$

True positives are all the relevant results within the subset (defined by k), and false positives are irrelevant. Given a dataset of 10 images, say we want to evaluate the top 5 results (k=5). Three of these are relevant to the query, while two are not. Precision for this system would be

$$Precision@5 = \frac{3}{3+2} = 3/5$$

Recall evaluates the relevancy of returned results based on all the relevant results present in the database. Its formula is

$$Recall@k = \frac{TruePositive@k}{TruePositives@k + FalseNegatives@k}$$

The false negatives in the formula depict all the relevant items that were not part of the final result set. Continuing our example from before, if we have four relevant results in the remainder of the dataset, its recall would be

$$Recall@5 = \frac{3}{3+4} = 3/7$$

Developers often face the precision-recall trade-off where they have to find a balance.

Both metrics are contrasting as precision showcases the system based on the true labels retrieved while recall judges the true labels left behind. An effective IR system must display reasonable values of both.

Advanced Metrics: NDCG and MAP

Advanced information retrieval metrics, such as NDCG (Normalized Discounted Cumulative Gain) and MAP (Mean Average Precision), are crucial for evaluating the performance of IR systems. These metrics are order-aware, meaning the value they produce is influenced by the order in which retrieved items are presented.

MAP (Mean Average Precision)

MAP builds upon the precision metric but enhances it by considering multiple values of k rather than just one. Here's how it works:

- Average Precision (AP): For a given k (e.g., 5), MAP calculates precision at each rank from 1 to k. It averages these precision values to obtain the Average Precision for that query.

- Multiple Queries: A robust IR system should handle a variety of user inputs. MAP computes the AP across several queries and then averages these results to provide a final MAP value, reflecting system performance across different scenarios.

This multi-query approach gives a better overall picture of how well the IR system performs in real-world conditions.

NDCG (Normalized Discounted Cumulative Gain)

NDCG relies on a ground truth ranking associated with each item in the database. Here's how it operates:

- User-defined Ranks: For instance, with a query like "White sports car with a red spoiler," images are ranked based on relevance. A perfect match might receive a rank of 5, while irrelevant images get a rank of 1, with partial matches falling in between.

- Calculating NDCG: The sum of the ranks of retrieved images is calculated, favoring more relevant results. However, to address the order of these results, a log-based penalty is introduced. This penalty adjusts scores based on the rank position, lowering scores for relevant images that appear lower in the result set.

- Normalization: Since NDCG scores can have no upper bound, normalization is crucial. The score is divided by the ideal score (where all relevant items are ranked at the top), resulting in a normalized score between 0 and 1.

Applying Metrics to Evaluate Systems:

IR metrics are vital for pre-deployment evaluations, helping developers fine-tune the search framework. Typically, a testing framework uses predefined queries and labeled documents to automate the evaluation of the IR system, highlighting queries that perform poorly.

Search engines like Google and Bing are prime examples of IR systems that benefit from these metrics. They utilize statistical algorithms and vector databases to sift through billions of documents, and evaluation metrics enhance search relevance and user satisfaction.

Conclusion reached :

Information retrieval metrics assess how well statistical algorithms retrieve documents in response to user queries. These metrics can be categorized into two groups:

- Order-unaware metrics: Such as Precision, Recall, and F1-Score, focus solely on the overall relevance of results, ignoring their order.

- Order-aware metrics: Like MAP and NDCG, these more advanced metrics evaluate performance based on multiple search scenarios and penalize the system for poor ordering of relevant results.

By leveraging these metrics, IR systems can continually improve their search relevance and provide a better user experience. Some enterprises also employ more complicated retrieval and evaluation frameworks to create robust search mechanisms.