Understanding Ranking and Probability in Information Retrieval

This document provides a structured explanation of ranking and probability concepts in Information Retrieval, as depicted in the provided slides. Each topic is explained step by step with examples for better comprehension.

Slide 1: Deriving a Ranking Function for Query Terms

In this section, we analyze the ranking function for query terms using probabilities. The key components include:

- Document frequency (df_t): Number of documents containing the term t.
- Relevant documents (S) and non-relevant documents (N - S).

We calculate the probability of relevance based on the presence and absence of terms in relevant and non-relevant documents.

Slide 2: Odds and Log Odds in Document Ranking

Here, the odds ratio is introduced to compare the likelihood of term appearance in relevant versus non-relevant documents. The formula:

Odds Ratio = $p_t / (1 - p_t)$: Probability of relevance / Probability of irrelevance.

Log odds simplify the calculation and are used as weights (c_t) for ranking.

Slide 3: RSV - Retrieval Status Value

RSV is the cumulative weight (log odds) of query terms found in a document. It is calculated as:

RSV_d = Sum(log(c_t)) for all query terms.

This provides a ranking score for documents relative to a query.

Example for Understanding Log Odds

Consider a query term 'machine' appearing in:

- Relevant documents: 10/50 (20%).

- Non-relevant documents: 5/50 (10%).

Log odds ratio = log((0.2 / 0.8) / (0.1 / 0.9)) = 0.58.

This positive value indicates higher relevance for documents containing 'machine'.

Slide 4: Smoothing and Avoiding Zero Probabilities

In practice, zero probabilities (e.g., if a term is missing in relevant documents) are avoided using smoothing techniques. These methods adjust probabilities to ensure calculations remain valid.