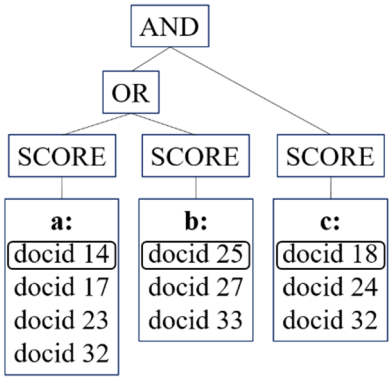
Sample Exam Questions

Feb 28, 2024

1. A document-at-a-time (DAAT) algorithm finds the next matching document id (docid) by repeatedly advancing docid iterators until a match is found. In the example below, the docid iterators are initialized to the first docid in each inverted list. Describe how the iterators <u>in this example</u> are moved to find the first matching document id (e.g., "Advance term t1 to docid d1. Advance t2 to docid d2."). Your answer must be clear about the order in which iterators are advanced and which docids they point to at each step.



Answer

- Advance a to at least 18 (to 23).
  - a → 17 → 23 is okay
- Advance c to at least 23 (to 24).
- Advance a to at least 24 (to 32).
- Advance c to at least 25 (to 32).
- Advance b to at least 32 (to 33).

2. Suppose that you are developing a new test collection using the Cranfield methodology. You plan to pool the results of four retrieval algorithms. You have six choices:
    - Unranked Boolean
    - Ranked Boolean
    - VSM using lnc.ltc,
    - BM25
    - Query likelihood
    - Kullbeck-Leibler Divergence

    Which four algorithms would you use to create a pool that covers as many relevant documents as possible? Justify your answer.

    Answer
    The best choices are Ranked Boolean, VSM using lnc.ltc, BM25, and query likelihood or KL divergence. Unranked Boolean is a poor choice because it won't find many relevant documents. It does not make sense to use both query likelihood and KL divergence because they are the same algorithm.

3. Language modeling

   a. Briefly describe what statistical language models are. **[4 points]**

   Answer

   A statistical language model is a probability distribution over word occurrences (unigrams) or word sequences (ngrams). The language model describes the probability of observing a term or sequence of terms in a sample of text.

   b. How does the query likelihood retrieval model use statistical language models to rank documents? **[6 points]**

   Answer

   The query likelihood retrieval model estimates the probability that a particular sample of text (the document) was produced by a particular language model (estimated from the query). This is a difficult problem because the query language model is very sparse (covers few terms) and has granular probabilities (because they are based on little frequency information), so Bayes Rule is used to transform it into an easier problem: Estimating the probability that a particular sample of text (the query) was produced by a particular language model (estimated from the document, and smoothed by the corpus language model). Typically all of the available language models are considered equally probable (i.e., uniform p(d)).

   c. Identify two reasons that smoothing is used in language modeling retrieval methods. Briefly justify your answer. **[4 points]**

   Answer

   Reason 1: To provide probability estimates for terms that are missing from the language model.

   Reason 2: To give greater importance to terms that are rare in the corpus and lower importance to terms that are frequent in the corpus (an idf-like effect).