Document Ranking Approaches for Query Relevance

Document Ranking Approaches for Query Relevance

1. Introduction

In information retrieval, accurately ranking documents based on user queries is essential to deliver relevant information efficiently.

In this report, we compare three models applied for ranking: the `all-MiniLM-L6-v2` model, Latent Semantic Indexing (LSI),

and a probabilistic model with iterative refinement. We examine each model's approach, formulas, and impact on results.

Document Ranking Approaches for Query Relevance

2.A all-MiniLM-L6-v2 Model

Model: all-MiniLM-L6-v2

Definition: A transformer-based language model that generates embeddings for text.
It captures semantic similarities in high-dimensional space. Each document and query are encoded into vectors.

Purpose: Embeddings represent the text in a way that captures context and meaning, allowing cosine similarity to rank documents.

Application: Documents are ranked based on cosine similarity between the encoded document and query vectors.

Document Ranking Approaches for Query Relevance

2.B Latent Semantic Indexing (LSI)

Model: Latent Semantic Indexing (LSI)

Definition: A dimensionality reduction technique using Singular Value Decomposition (SVD) to capture latent relationships between terms and documents.

Formula: SVD decomposes the TF-IDF matrix A into $A = U \Sigma V^T$, where:

- U: Left singular vectors (document-topic relationships)
- Σ: Singular values (importance of each topic)
- V^T: Right singular vectors (topic-term relationships)

Purpose: Reduces the vocabulary space by retaining only the top singular values, capturing the most significant patterns.

Application: Documents and queries are transformed into a lower-dimensional space, with similarity measured in this latent space.

Document Ranking Approaches for Query Relevance

2.C Probabilistic Model with Iterative Refinement

Model: Probabilistic Model with Iterative Refinement

Definition: A ranking model that uses probabilities
to weigh the relevance of terms based on their occurrence in relevant and non-relevant documents.

Relevance Formula:

- $P(t_i | R) = \frac{V_i + 0.5}{V + 1}$
- $P(t_i | NR) = \frac{n_i - V_i + 0.5}{N - V + 1}$

Purpose: Iteratively refines term relevance by updating probabilities based on a selected subset of ranked documents.

Application: For each query, calculates similarity using a weighted log-odds formula. Updates probabilities to enhance accuracy with each query.

3. Comparison of Methods

Model	Characteristics	Pros	Cons
all-MiniLM-L6-v2	Transformer-based embeddings capturing semantics	Captures deep context, adaptable for similarity ranking	Requires GPU for speed, embeddings can be high-dimensional
LSI	Latent space, reduces noise	Captures term dependencies, effective for large corpora	Sensitive to parameter choice, may lose term specificity
Probabilistic Model	Iterative probability-based relevance scoring	Dynamically improves with queries, considers relevance	Computationally intensive, sensitive to initial scores

4. Influence on Results

Influence on Results:

- **all-MiniLM-L6-v2**: Provides embeddings that capture semantic relationships in context, allowing for more accurate similarity rankings in high-dimensional space.
  However, embeddings are computationally intensive and best suited for GPU environments.

- **LSI**: Reduces dimensionality by capturing latent topics, improving results when terms are not direct matches but semantically related.
  May introduce noise if too many dimensions are retained.

- **Probabilistic Model**: Adapts dynamically, refining accuracy over time as relevance probabilities are adjusted based on feedback from prior queries.
  This model is advantageous in evolving search scenarios but requires additional computation.

5. Conclusion

Each model offers unique strengths: all-MiniLM-L6-v2 is best for capturing deep semantic similarity, LSI aids in capturing latent meaning, and the probabilistic model provides a dynamically refined ranking approach. Combining these methods may yield optimal relevance and adaptability in complex and evolving corpora.