Enhancing Information Retrieval Using Topic Modelling

Kavisha Madani, Vishaka Nair, Srushti Bhagchandani and Shubham Gupta

1. Problem Statement

The project aims at improving the efficiency and effectiveness of information retrieval systems by merging advanced techniques like CUR decomposition and Latent Dirichlet Allocation (LDA). This will retrieve the document that best matches a given user query by utilizing dimensionality reduction and topic modeling to align the topic distributions of both queries and documents. It is focused upon the issues of scalability, relevance ranking, and semantic understanding integrated into large-scale retrieval tasks.

2. Dataset

- **MS MARCO Dataset:**
  - Contains millions of real-world query-document pairs.
  - Designed for training and testing information retrieval systems.
  - Each query is associated with a relevant document.
- **Reuters Dataset:**
  - A collection of news articles used for topic modeling and text classification.
  - Covers a variety of topics and supports robust extractions using LDA.
- **BD News Dataset:**
  - Filtered news dataset from Bangladesh, spanning politics, economy, and culture.
  - LDA applied to evaluate generalization across various text corpora when integrated with CUR.
  - Compared performance with the Gensim library's LDA function.

2.1. Preprocessing Steps

- Tokenization.
- Removal of stopwords.
- Normalization of text.
- Vectorizing query-document pairs for CUR decomposition and topic modeling.

3. Evaluation Metrics

3.1. MS MARCO Dataset

- **Precision@K:** Proportion of relevant documents retrieved in the top-K results.
- **Recall@K:** Proportion of relevant documents retrieved out of all relevant documents.

Introduction to Information Retrieval (IT550), Autumn 2024, DAIICT, India

202318007 (K. Madani); 202318041 (V. Nair); 202318047 (S. Bhagchandani); 202318052 (S. Gupta)

3.2. All Datasets (MS MARCO, Reuters, BD News)

- **Perplexity:** Lower score indicates better generalization capability.
- **Coherence Score:** Higher score reflects better semantic interpretability and topic quality.

4. Process Overview

The project begins by importing the text corpus from three diverse datasets: Reuters, BD news, and MS MARCO. The goal is to ensure the datasets are sufficiently large and varied to represent different topics and facilitate comprehensive evaluation of the retrieval mechanism. To prepare the data, text preprocessing is performed. This involves tokenization, stemming, lemmatization, lowercasing and stop word removal.

After preprocessing, the system is designed to accept a query input from the user. This query is written in natural language and serves as the search key to retrieve relevant documents from the dataset. To optimize the system's performance, CUR decomposition is applied to the term-document matrix, which represents the frequency of terms across documents. CUR decomposition selects a subset of rows (documents) and columns (terms) based on their importance and outputs three components: the C matrix (subset of columns), the U matrix (interaction information between rows and columns), and the R matrix (subset of rows). This low-rank approximation reduces the dimensionality of the matrix while retaining critical term-document relationships, making subsequent computations faster and more efficient.

The CUR-reduced term-document matrix is then fed into the Latent Dirichlet Allocation (LDA) algorithm. Instead of using the entire matrix, only the C matrix is used, which significantly reduces the computational load. LDA identifies two key distributions: document-topic distributions (probabilities of each topic in a document) and topic-word distributions (probabilities of each word in a topic). By working on this reduced matrix, the vocabulary space is compressed without losing meaningful patterns, enabling faster topic extraction.

To enhance the retrieval process, dynamic query expansion is implemented. This technique enriches the user's query by adding synonymous or related terms to corm the expanded appropriate weights are assigned to balance their contributions. The expanded query captures the user's intent more effectively.

The system offers two methods for retrieving information: sentence transformers and topic similarity.

1. **Sentence Transformers**: Compute cosine similarity between query and document embeddings for precise results, albeit with higher computational cost.
2. **Topic Similarity**: The cosine similarity between the query's topic distribution and each document's topic distribution is computed, allowing the system to sort and retrieve the most relevant documents. This method is computationally efficient and provides effective results.

This is how CUR reduces dimensionality, enabling faster computations, while LDA and query expansion ensure that retrieved documents align contextually with the user's input. This end-to-end process enhances retrieval accuracy and system performance.

5. Results

5.1. Hyperparameter Tuning

- LDA hyperparameters (*topics k*, *alpha*, *beta*) were optimized using perplexity and coherence scores.
- Observed trade-off between the number of topics and coherence score:
  - Optimal topics: 12 (Reuters), 8 (BD News).
  - Larger topic numbers increased perplexity due to noise sensitivity.
- Systematic tuning improved semantic quality of topics and overall retrieval system performance.

5.2. Dataset-Specific Results

- **MS MARCO Dataset:**
  - Precision@5: 0.5970 (10k) → 0.6742 (30k).
  - Recall@5: 0.6818 (10k) → 0.7742 (30k).
  - Coherence: -5.4797 (10k) → -10.130 (30k).
  - Perplexity: 7770.9585 (10k) → 1365.28 (30k).
- **Reuters Dataset:**
  - Perplexity: 1093.6478 (CUR-LDA) vs. 3117.588 (Gensim Standard) and 2193.6 (Gensim TF-IDF).
  - Coherence: -2.657 (CUR-LDA) vs. -1.8551 (Gensim Standard) and -12.7924 (Gensim TF-IDF).
- **BD News Dataset:**
  - Perplexity: 594.5621.
  - Coherence: 166.8932.

6. Key Challenges and Learnings

- **CUR Decomposition**:
  - Reduces dimensionality while retaining key features, addressing data sparsity in term-document and tf-idf matrices.
  - Extends beyond SVD by focusing on extracting important features, enhancing topic interpretability.
  - Acts as a regularizer for LDA, reducing noise and preventing overfitting.
- **LDA Topic Modeling**:
  - Improved semantic understanding of queries and documents in the reduced feature space.
  - Involved iterative tuning of parameters (e.g., number of topics, alpha, beta) for optimal results.
  - Combined with CUR, ensures compatibility in reduced-dimensional space for effective topic distributions.
- **Challenges**:
  - Computational intensity of CUR for large datasets.
  - Mathematical and implementation hurdles to align CUR-reduced features with topic distributions.
  - Robust evaluation with real-world datasets like MS MARCO highlighted the need for comprehensive performance metrics.

7. References

1. David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, null (3/1/2003), 993–1022.
2. M.W. Mahoney, P. Drineas, CUR matrix decompositions for improved data analysis, Proc. Natl. Acad. Sci. U.S.A. 106 (3) 697-702, https://doi.org/10.1073/pnas.0803205106 (2009).