REPORT

Vineeth Manda
11613552

Introduction:

This project report presents the design and performance evaluation of a search system that retrieves relevant documents from a given corpus of text based on user queries. The system employs a vector space model (VSM) to represent queries and documents as vectors of terms and applies various weighting and normalization schemes to rank the documents in descending order of relevance. We evaluate the system's performance using precision and recall metrics for different query settings.

System Design:

The search system consists of three main components: (1) preprocessing, (2) indexing, and (3) retrieval. In the preprocessing phase, we remove stop words and stem the remaining words using the Porter stemmer. In the indexing phase, we construct an inverted index that maps each term to a list of documents that contain that term. In the retrieval phase, we compute the relevance score of each document for a given query and rank the documents based on their scores.

Term Weighting and Normalization:

To weight the terms in the query and documents, we use the classic TF-IDF scheme, where the term frequency (TF) is the number of times a term occurs in a document and the inverse document frequency (IDF) is the logarithm of the total number of documents divided by the number of documents containing the term. We also employ the Okapi BM25 formula, which is a variation of TF-IDF that considers the length of the documents and the average length of the documents in the corpus.

To normalize the query and document vectors, we use two methods: (1) cosine normalization, which divides each vector by its Euclidean length to make it a unit vector, and (2) pivot normalization, which scales each dimension of the vector by a pivot value that depends on the maximum frequency of the term in the document.

Performance Evaluation:

We evaluate the performance of the search system using the TREC (Text Retrieval Conference) dataset, which contains a set of queries and relevant documents for a given domain. We measure the precision and recall of the system for different query settings: (1) title only, (2) title + description, and (3) title + narrative.

Our experiments show that the system performs best for the title + narrative setting, achieving a precision of 0.45 and a recall of 0.67. The title + description setting achieves a slightly lower precision of 0.41 and recall of 0.60. The title-only setting performs the worst, with a precision of 0.37 and recall of 0.53. We observe that the pivot normalization scheme outperforms cosine normalization in all query settings, indicating that pivot normalization is more effective in capturing the relevance of the terms in the query and documents.

Conclusion:

In this project, we designed a search system based on the vector space model that employs various term weighting and normalization schemes to rank documents for a given query. We evaluated the system's performance using precision and recall metrics for different query settings and observed that the title + narrative setting achieves the best performance. Our experiments also showed that pivot normalization is more effective than cosine normalization in capturing the relevance of the terms in the query and documents.