IR model
Boolean Model

Information Retrieval

Chung-Ang University

CAU 2023

1st Semester

Document:

Project Report

(Version 1.0)

(Date: 04-06-2023)

Table of contents

Introduction

A – Background and Objectives

Information retrieval (IR) is a crucial task in the field of natural language processing. The goal of IR is to retrieve relevant information from a large corpus of documents in response to a user's query.

In this project, we implemented a Boolean model for IR, which is a classic retrieval model that uses Boolean operators (AND, OR, NOT) to combine query terms and document terms.

The objective of this project is to develop an IR system using the Boolean model that is capable of accurately retrieving relevant documents in response to user queries.

B – Problem Statement

In this project, we are given a text collection of reviews and four sample queries ("korean", "squid game", "vip", and "life").

Our task is to implement a Boolean model for IR, which will allow us to retrieve relevant documents from the text collection in response to the queries. Additionally, we need to evaluate the performance of our IR system using standard evaluation metrics and provide an analysis of the results.

C – Scope and Limitations

The scope of this project is limited to the implementation of a Boolean model for IR and its evaluation using the provided text collection and queries. We will not be exploring other IR models or using other datasets for testing.

Additionally, the evaluation metrics used in this project will only provide a basic evaluation of the IR system and may not reflect its performance in all contexts.

Developer

In this page you will find the name, as well as the Student ID of the developer of this project:

1. **Yeray Cordero Carrasco**

- Student ID: 50221570
- Name in Korean: 예라이 코데로 카라스코

Methodology

A – Data Collection

The text collection used in this project consists of reviews that were downloaded given by the professor. The collection contains a total of 264 reviews in TXT format.

B – Data Pre-processing

The first step in pre-processing the data involved tokenization, where the text in each document was split into individual words.

C – Indexing

The next step was to create an inverted index, which is a data structure that stores the terms in the corpus and the documents that contain each term.

This was done by first creating a dictionary of terms and their corresponding posting lists, which contains the document IDs where the term appears. Each posting list was sorted by document ID to allow for efficient intersection and union operations during query processing.

D – Query Processing

Queries were processed using the Boolean model to combine query terms and document terms. The query was first pre-processed using the same steps as the documents, and then the terms in the query were used to retrieve the relevant documents from the inverted index.

E – Evaluation Metrics

The performance of the IR system was evaluated using a standard evaluation metric: precision. Precision measures the proportion of retrieved documents that are relevant.

Results and discussion

In this project, we implemented a Boolean Model Information Retrieval system based on the specifications provided. The system was implemented in Java using an inverted index data structure to store the documents and terms.

To evaluate the system's performance, we used precision as the evaluation metric. Precision measures the proportion of relevant documents among the total retrieved documents. We manually assessed the relevance of documents based on whether they contain the exact query term(s).

We tested the system with four queries: "korean", "squid game", "vip", and "life". The evaluation results for each query are presented in the table below.

Query	***Relevant Documents***	***Retrieved Documents***	***Precision***
korean	66	66	**1.0**
squid game	56	56	**1.0**
vip	7	7	**1.0**
life	24	24	**1.0**

As we can see from the table, the system achieved perfect precision for all queries, meaning that all the retrieved documents were relevant to the query. However, it is important to note that this evaluation metric only considers the number of relevant documents retrieved, and not the total number of relevant documents in the corpus. In a real-world scenario, there may be more relevant documents that were not retrieved by the system.

The Boolean Model is a simple but powerful model for Information Retrieval. It allows for exact match searching and is easy to implement. However, it suffers from some limitations, such as not being able to handle synonymy (different terms with the same meaning) or polysemy (the same term with different meanings). Additionally, the Boolean Model does not consider the relevance of documents based on the query context.

In conclusion, we have successfully implemented a Boolean Model Information Retrieval system and evaluated its performance using precision as the evaluation metric. While the system achieved perfect precision for all queries, it is important to note the limitations of the Boolean Model and consider other evaluation metrics and models for more realistic Information Retrieval scenarios.

Conclusion and Recommendations

A – Conclusion

In this project, we have implemented a Boolean Model for Information Retrieval and tested it on a text collection containing reviews. We have evaluated the performance of the system using the precision evaluation method.

The results show that the Boolean Model can effectively retrieve relevant documents for simple queries with high precision. However, it may not be suitable for more complex queries as it does not take into account the relevance ranking of the documents.

In addition, we have identified several limitations of the current implementation, such as the lack of stemming and stop-word removal, which may negatively impact the system's performance. Nevertheless, the implementation can be improved by incorporating these techniques.

B – Recommendations for Future Work

Based on the results and limitations of the current implementation, we recommend the following improvements:

- Incorporate stemming and stop-word removal techniques to improve the system's performance.
- Implement additional evaluation methods such as recall and F1-score to better evaluate the system's performance.
- Experiment with other Information Retrieval models such as Vector Space Model and BM25 to compare and evaluate their performance.
- Increase the size of the text collection and test the system on a larger dataset to evaluate its scalability.

Overall, the implementation of the Boolean Model provides a good foundation for Information Retrieval, and with further improvements, it can be a useful tool for searching large collections of text data.

Appendices

A – Source Code

The source code for our Boolean retrieval model is available on GitHub at the following link: https://github.com/yeray142/IR_model

I will also upload all the relevant files to the e-class.

B – Evaluation Metrics Calculation

The precision of the Boolean model was calculated as the ratio of the number of relevant documents retrieved to the total number of documents retrieved for each query.

The precision for each query was found to be 1.0, indicating that all retrieved documents were relevant.

However, precision alone may not be sufficient to fully evaluate the performance of an IR system, and other metrics such as recall and F1 score may also be considered.