Information Retrieval

Academic Year 2025-2026

Course Project Progress Report

Xuanlin Chen

1 Design

The system employs a hybrid search architecture that combines traditional keyword-based search Apache Solr with deep learning-based semantic search LaBSE. The hybrid approach balances retrieval speed with accuracy, ensuring both precise matches and semantically relevant results are returned.

The system indexes multiple fields including title, menu item, content, ingredients, menu category, and store name. Multi-field indexing enables flexible multi-dimensional search, allowing users to query from different perspectives. Field weighting prioritizes store names and menu items over general titles, reflecting business logic where specific menu items and store locations are more important for user queries.

2 Search Engine Implementation

2.1 Solr Keyword Search Engine

Apache Solr serves as the core keyword search engine with the following configuration:

- **Text Analysis Pipeline**: Standard tokenization, lowercasing, stemming, synonym expansion, edge n-gram filtering for prefix matching, and stop word removal
- **Query Parser**: Extended Dismax parser supporting multi-field search, phrase matching, and boost queries

2.2 LaBSE Semantic Search Engine

The system integrates LaBSE model for semantic understanding:

- **Embedding Generation**: Maps queries and documents into a 768-dimensional semantic space
- **Similarity Computation**: Uses cosine similarity to measure semantic relatedness between query and document embeddings

3 Search Interface Implementation

3.1 Query Processing and Execution

The search interface processes user queries through multiple stages:

- **Query Input**: Accepts natural language queries from users
- **Parallel Execution**: Simultaneously executes keyword search via Solr and semantic search via LaBSE API