A Cross-lingual Search Engine for Retrieval of Green House Gas Emission Factor

Yachai Limpiyakorn1, Nattapot Nuwong2

1, 2 Department of Computer Engineering, Chulalongkorn University, Bangkok 10330, Thailand

Corresponding Author: Email address: yachai.l@chula.ac.th

*Abstract*—*Climate change is a critical global challenge primarily driven by greenhouse gas (GHG) emissions. Managing GHG data effectively is essential for accurate carbon footprint assessments. However, existing information retrieval systems often lack cross-lingual search often lack cross-lingual search capabilities, limiting access to Green House Gas Emission Factor (EF) data across languages, This study present a cross-lingual search engine designed to retrieve EF data effectively, integrating dictionary-based synonym matching and Elasticsearch indexing, The system enables seamless bilingual search functionality, it supports synonym-based search expansion, ensuring relevant result regardless of query language. By implementing an automated pipeline for data retrieval and indexing, the system enhances retrieval accuracy and efficiency thereby aiding organizations in precise carbon footprint calculations.*

Keywords— Cross-Lingual, Information Retrieval, GHGs Emission Factor, Synonym.

# Introduction

The world is currently facing extreme weather conditions, primarily due to the continuous and substantial emissions of greenhouse gases (GHGs) into the atmosphere. These emissions have widespread impacts on ecosystems, climate, biodiversity, and the quality of life at both local and global levels. Consequently, the effective monitoring and management of GHG emissions data is of paramount importance, especially for Thailand. The principal agency responsible for assessing, collecting, compiling, and disseminating GHG-related information is the Thailand Greenhouse Gas Management Organization (TGO), which operates under the Ministry of Natural Resources and Environment. TGO has developed comprehensive databases and report covering a wide range of sectors including industry, energy, agriculture, and communities as well as information for assessing and calculating carbon footprints. Moreover, it determines GHG emissions factors (EF) based on international standards and guidelines provided by the Intergovernmental Panel on Climate Change(IPCC). However, despite the TGO’s efforts to publish GHG emission factor data in various formats on its website (such as online databases and electronic files), there remains a gap the absence of a search system that can comprehensively handle EF queries in a cross-lingual context. In the content provide on TGO’s Website, some EF keywords appear in English, some in Thai, and some in both languages. Additionally, user preferences vary, with some users favoring Thai search terms while others prefer English. These linguistic discrepancies can cause conventional single-language search approaches to yield inaccurate or incomplete result. This research proposes the development of a cross-lingual search program for retrieving of GHG emission factor. The system is designed to support in both Thai and English and to incorporate synonym matching to enhance user accessibility regardless of language. It enables rapid and comprehensive searches, with the search result displaying EF entries across different versions to allow users to select the most appropriate value. This functionality is particularly beneficial for practitioners who need to choose correct EF values when preparing an organization’s carbon footprint report, thereby reducing errors, avoiding redundant work, and ensuring efficient resource utilization.

# Background

1. *EF (Greenhouse gas emission Factor)*

The Greenhouse gas Emission Factor is calculated based on the amount of greenhouse gas emitted and absorbed per unit of activity. It is used to assess the volume of greenhouse gas emission resulting from various activities, such as energy consumption, fuel consumption, transportation, or industrial production processes. This factor is a critical variable that enables organizations, government agencies, small businesses to calculate and analyze EF. It is essential for reporting, impact monitoring, and planning strategies to reduce carbon footprint. The calculation relies on secondary data sources, ranked by reliability from highest to lowest [1-3]

1. Thailand Greenhouse Gas Management Organization (TGO)
2. Research studies conducted in Thailand
3. Publicly available databases, including Open LCA
4. Intergovernmental Panel on Climate Change (IPCC)
5. *IP (Information Retrieval)*

Information Retrieval is a fundamental concept in information science that focuses on retrieving relevant documents from a large collection based on user queries. Various model have been developed to improve retrieval efficiency and effectiveness, including

1. Fundamental IR Models

• *Boolean Model*: Uses logical operators such as AND,OR and NOT to define search conditions. Documents that meet all conditions are retrieved based on an exact match, but this model lacks ranking mechanisms to order results by relevance.

• *Vector Space Model (VSM)*: Represents documents and queries as vectors in a high-dimensional term space. The relevance of a document to a query is computed using cosine similarity, allowing for ranked retrieval based on degree of similarity.

*• Probabilistic Models (e.g BM25*): Estimate the probability that a document is relevant to a query. These models incorporate factors such as Term Frequency (TF), Inverse Document Frequency (IDF), and document length normalization to refine retrieval rankings

1. *CLIR (Cross-Lingual Information Retrieval)[4]*

Traditional information Retrieval systems assume that queries and documents are in the same language. However, in CILR users may input queries in one language (e.g., Thai) while expecting relevant documents in another language (e.g., English) or in both languages. To enable efficient CLIR, a mechanism is required to bridge linguistic differences between languages[5, 6]. Two primary approaches are widely employed

1. Synonym-Based (Dictionary-Based)[7]

This approach relies on a bilingual dictionary containing predefined word-pair mappings between Thai and English. During the tokenization and indexing process, query expansion is performed to include synonymous terms in the target language. For example, if a user searches for “ก๊าซเรือนกระจก” the system will retrieve documents containing “greenhouse” and “gas greenhouse” as well. The bilingual dictionary operates as part of domain-specific knowledge and includes the following components

* + *Alternative Labels:* Terms that can be used interchangeably and the same meaning such as abbreviations, acronyms, or spelling variations.

*example:* CTO – Chief Technology Officer

* + *Synonyms:* Words or phrase that convey the same or similar meaning despite minor contextual differences.

*example:* human - homo sapiens, mankind

* + *Taxonomy:* A hierarchical system that organizes terms into structured categories and defines relationships among them.
  + *Ontology:* A formal representation of complex relationships between entities in a domain, including hierarchical dependencies (e.g. employee reporting to a manager)
  + *Knowledge Graph:* A practical implementation of ontologies that explicitly defines entity relationships  
    example: Michael is an employee, Micheal report to Jim, thus Jim is Micheal manager

2. Embedding-Based

This approach leverages Multilingual Natural Language Processing (NLP) models, such as BERT, LaBSE and others, to encode both Thai and English sentences into a shared vector space (latent space). When a user inputs query in Thai, the system converts the query into a vector and matches it against document vectors, regardless of their language. If two phrases have similar meaning, their vector representations will be close to each other in the latent space.

1. *NLP (Natural Language Processing)[8]*

Natural Language Processing is a branch of Artificial Intelligence (AI) that focuses on enabling computers to understand, interpret, and process human language effectively. In the context of Cross-Lingual Information Retrieval (CLIR), NLP plays a crucial role in data preprocessing and enhancing the understanding of the languages used in search queries. This study specifically focuses on Thai and English, which exhibit significant differences in linguistic structure, word segmentation, and specialized vocabulary processing.

* + *Tokenization:* Thai does not use spaces between words, making word segmentation a challenging task. Specific tools as ICU Tokenizer in Elasticsearch are employed to segment words appropriately. In contrast, English typically uses the Standard Tokenizer, with additional steps such as stemming or lemmatization to normalize words and improve search accuracy.
  + *Stop Words:* The identification of stop word, such as and, is, the helps reduce noise and improve search efficiency. Additionally, the use of synonyms in both Thai and English can enhance retrieval effectiveness, particularly in Dictionary-based CLIR system.
  + *NER (Named Entity Recognition):* In certain cases, it is essential to identify named entities or domain-specific terminology, such as chemical names. This can be achieved using NER to improve entity matching and retrieval performance*.*

1. *Technology stack*

• *Elasticsearch[9]:* Elasticsearch is a distributed search engine that supports full-text search, structured search, and vector search. It provides plugins and analyzers for Thai language processing, such as the ICU Tokenizer, and allows the configuration of Synonym Filter for cross-lingual search.

1. Full-text: Supports text search in both Thai and English by configuring a custom analyzer and synonym filter
2. Synonym Matching: Uses the synonym filter to match equivalent term across languages, such as “LPG”, “Liquefied Petroleum Gas” , “ก๊าซหุงต้ม”

• *FastAPI[10]:* FastAPI is a high-performance Python web framework designed for building REST APIs. It is lightweight and easy to use, facilitating seamless integration between frontend applications and Elasticsearch

• *Apache Airflow[11]:* workflow management platform used for designing, orchestrating, and monitoring Directed Acyclic Graphs (DAGs). It is particularly useful for managing data pipelines and ETL (Extract, Transform, Load) processes, ensuring efficient task execution and scheduling

# METHODOLOGY

Fig. 2 illustrates the research process, beginning with the preparation of EF data from Thailand Greenhouse gas Management Organization. Subsequently, data verification and anomaly handling are performed to ensure that data is in a suitable format. The search configuration is then defined, and the processed data is ingested into Elasticsearch

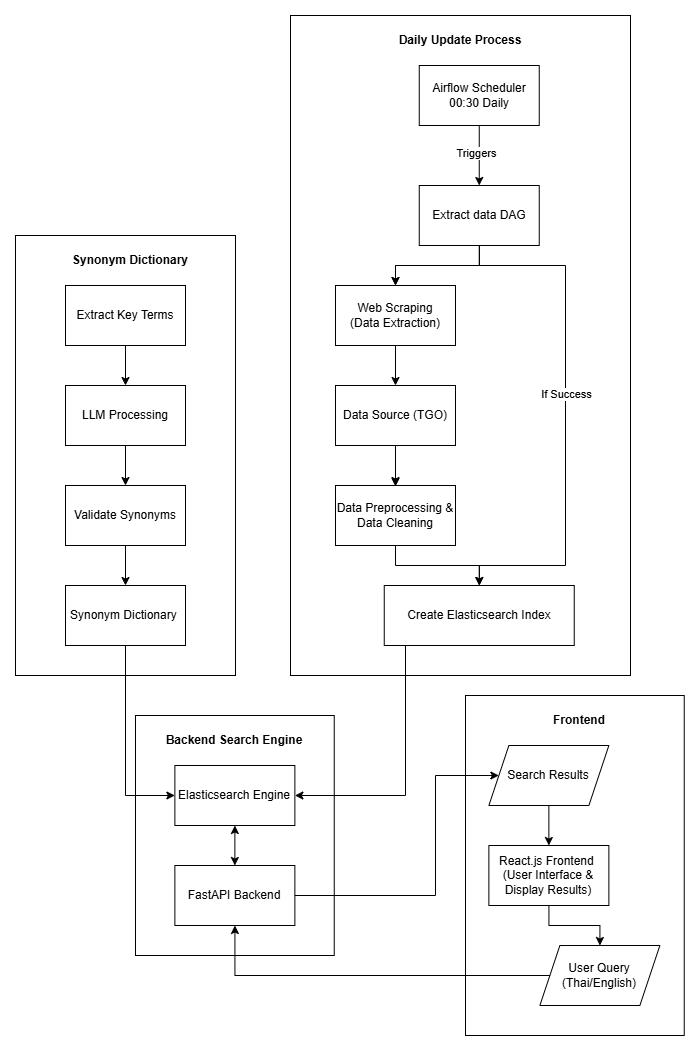


Fig. 2. Research methodology.

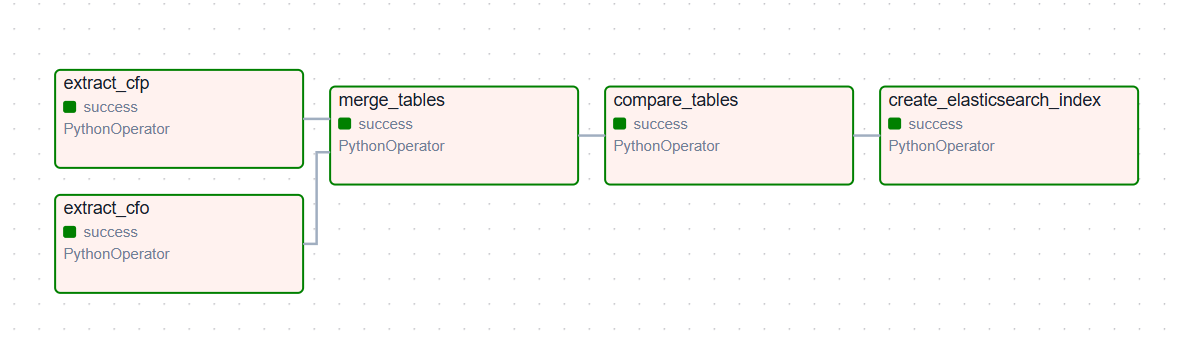


Fig. 3 Data pipeline workflow.

1. *Experiment*

*1. Data Preprocessing*

Airflow Scheduler plays a crucial role in the EF data collection process from the TGO website by managing and automating scheduled tasks. The system runs a processing pipeline to update data daily at 12:30 AM, as depicted in Fig. 3.

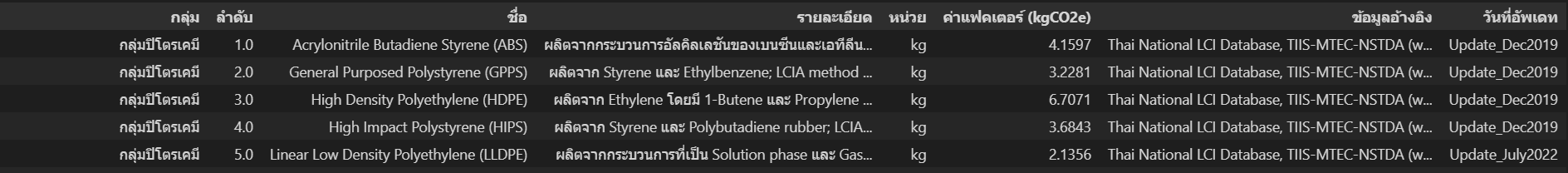


Fig. 4. Example Table of Greenhouse Gas (GHG) Emission Factors.

Fig. 4 presents an example table of Greenhouse Gas (GHG) Emission Factors, obtained from the Thailand Greenhouse Gas Management Organization (TGO). The dataset consists of multiple tables and various types of GHG emission factors, which require further processing. The data is stored in CSV format for subsequent computational analysis and processing*.*

1. *Index Creation in Elasticsearch*

The preprocessed and structured data is stored in an Elasticsearch index to support efficient and flexible search capabilities across both Thai and English languages. The search index is designed to enhance retrieval accuracy and consists of the following key components

*• Index Structure*: Organizes data into relevant categories, such as chemical names and measurement units

*• Language Analyzer Configuration*: Defines tokenization and word analysis methods for both Thai and English

*• Synonym Token Filter*: Enables synonym matching during search queries to enhance retrieval accuracy

The search index configuration includes filters and analyzers, as illustrated in Fig. 5



Fig. 5. Configuration Settings for Creating a Search Index.

1. Filters: Filters are used to process and refine tokens generated by the tokenizer to improve search effectiveness. The following filters are applied:

* thai\_english\_synonym\_filter: Handles synonym transformation between Thai and English terms. For instance,  
  *Anthracite* ↔ *แอนทราไซต์* ↔ *ถ่านหินแข็ง*.
* edge\_ngram\_filter: Supports autocomplete suggestions by generating potential search query completions.

2. Analyzers: Analyzers are text-processing pipelines consisting of a tokenizer and filters to optimize search performance. Examples include:

* thai\_synonym\_analyzer: Processes Thai text and integrates synonym recognition to improve search flexibility.

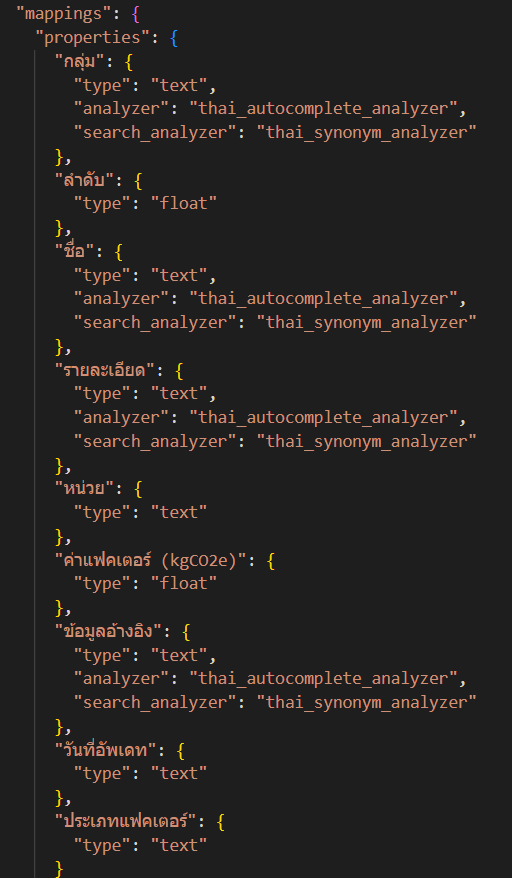


Fig. 6. Configuration Settings for Creating a Search Index.

Fig. 6 illustrates the document structure within the Elasticsearch index, specifying column names and data types. This configuration ensures that both analyzers and search analyzers correctly interpret the data. The “name” column utilizes both autocomplete\_index\_analyzer for search queries and thai\_synonym\_analyzer for synonym recognition, allowing for more comprehensive and accurate search results.

*3. Creating a Synonym Dictionary*

A synonym dictionary is constructed to facilitate cross-lingual search, ensuring that search queries return comprehensive results in both Thai and English. This process involves compiling words and phrases with equivalent meanings or interchangeable terms, including abbreviations and chemical symbols, all of which share the same semantic meaning. The synonym dictionary plays a crucial role in improving search accuracy and coverage. Examples include:

* *Liquefied Petroleum Gas*, *LPG*, *ก๊าซปิโตรเลียมเหลว*,   
  *ก๊าซหุงต้ม*
* *Anthracite*, *แอนทราไซต์*, *ถ่านหินแข็ง*

The process of generating this synonym set is performed using a Large Language Model (LLM) to identify and analyze relevant terminology. The workflow begins with inputting emission factor terminology into the LLM, which translates and aligns the meanings of Thai and English terms.

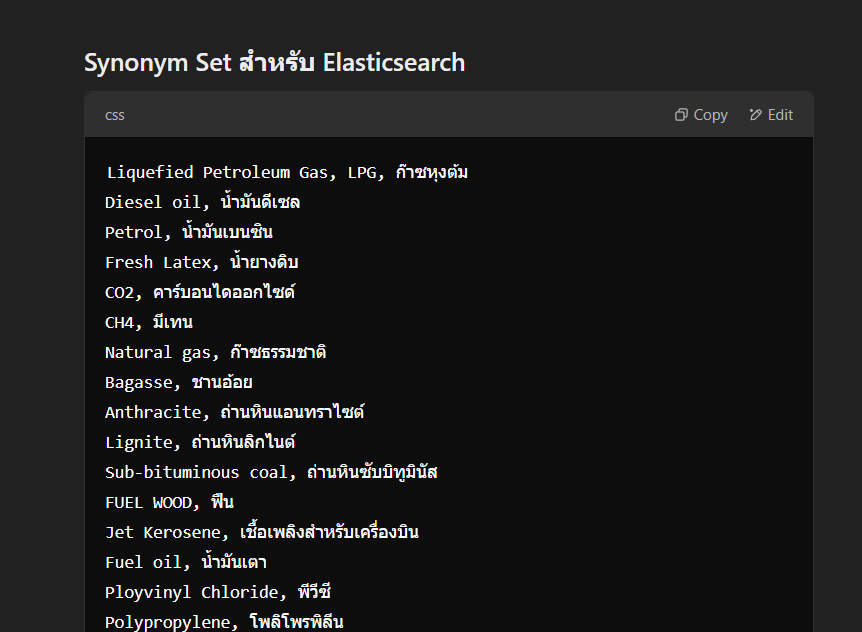


Fig. 7. Results Obtained from Executing the Command.

Fig. 7 presents the results generated from the executed command or prompt, which was formulated as follows

"Generate a Synonym Dictionary in the form of a Synonym Set for domain-specific terms that share the same meaning in both Thai and English, including abbreviations if applicable. Format the output in a structure compatible with Elasticsearch, such as ก๊าซหุงต้ม, LPG, Liquefied Petroleum Gas"

*4. Backend*

The backend is developed using FastAPI, serving as an intermediary between the frontend and Elasticsearch. Its primary function is user query processing, where it:

* Receives search queries from users (in either Thai or English) via an API.
* Forwards the queries to Elasticsearch and retrieves the results.
* Processes the retrieved data, ranks the results based on relevance, and returns them to the frontend

*5. Frontend*

The frontend is developed using React.js, providing an interactive search interface for users. It includes:

* A search box that supports both Thai and English queries.
* A structured results display, such as a table format, for clarity and ease of understanding.

1. *Evaluation*

Performance Evaluation

To assess the effectiveness of IR systems, several widely used evaluation metrics are applied

• *Precision:* Measures the proportion of retrieved documents that are relevant. It is calculated as



where TP (True Positives) represents relevant documents retrieved, and FP (False Positives) represents non-relevant documents retrieved.

• *Recall:* Evaluates the system ability to retrieve all relevant documents from dataset. It is calculated as



where FN (False Negatives) refers to relevant documents that were not retrieved.

• *Mean Average Precision (MAP):* A ranking-based evaluation metric that computes the average precision at different recall levels across all queries. It is particularly useful for systems that return ranked lists of results



where

AP(q) = represents the average precision for query q,

Q = the total number of queries.

These metrics are essential for assessing IR system effectiveness and optimizing retrieval algorithms to enhance user experience and search accuracy.

TABLE I. Experimental Result.

|  |  |
| --- | --- |
| Method | Result |
| Precision | X |
| Recall | Y |
| MAP | Z |

# CONCLUSION

This research presents the development of a cross-lingual search engine for retrieving Emission Factors, addressing the challenges posed by language barriers and synonym variations in emission factor datasets. The system integrates Elasticsearch, FastAPI, and React.js to enable efficient, multilingual, and scalable search functionalities. The core contributions of this study include

• Cross-lingual Information Retrieval (CLIR) The implementation of a synonym dictionary for Thai-English term matching ensures comprehensive search coverage, allowing users to retrieve relevant emission factor data regardless of the input language

• Efficient Indexing and Search Optimization Using Elasticsearch’s analyzers, filters, and tokenization techniques, the system enhances query precision, recall, and ranking effectiveness

• Automated Data Processing Pipeline The integration of Apache Airflow facilitates real-time data updates from Thailand Greenhouse Gas Management Organization (TGO), ensuring that emission factor data remains accurate and up to date

REFERENCES

[1] T. G. G. M. O. P. Organization). "Thai Carbon Label." <https://www.tgo.or.th/2023/index.php/th/> (accessed January 15, 2025.

[2] O. Nexus. "OpenLCA Nexus Databases." <https://nexus.openlca.org/databases> (accessed January 15, 2025.

[3] I. P. o. C. C. (IPCC). "IPCC Emission Factor Database (EFDB)." <https://www.ipcc-nggip.iges.or.jp/EFDB/main.php> (accessed 15 January 2025).

[4] G.-A. Levow, D. W. Oard, and P. Resnik, "Dictionary-based techniques for cross-language information retrieval," *Information processing & management,* vol. 41, no. 3, pp. 523-547, 2005.

[5] P. Nikesh, S. M. Idicula, and S. D. Peter, "English-Malayalam cross-lingual information retrieval—an experience," in *2008 IEEE International Conference on Electro/Information Technology*, 2008: IEEE, pp. 271-275.

[6] D. T. Nguyen and C. T. Nguyen, "Cross-Lingual Information Retrieval Model for Vietnamese-English Web Sites," in *2010 Second International Conference on Computer Modeling and Simulation*, 2010, vol. 4: IEEE, pp. 254-257.

[7] T. Grainger, D. Turnbull, and M. Irwin, *AI-Powered Search*. Simon and Schuster, 2025.

[8] S. Sun, C. Luo, and J. Chen, "A review of natural language processing techniques for opinion mining systems," *Information fusion,* vol. 36, pp. 10-25, 2017.

[9] M. Konda, *Elasticsearch in Action*. Simon and Schuster, 2024.

[10] S. Ramírez. "Modern, Fast (high-performance) web framework for building APIs with Python." <https://fastapi.tiangolo.com/> (accessed January 20, 2025, 2025).

[11] A. S. Foundation. "Apache Airflow: Platform to programmatically author workflows." <https://airflow.apache.org/> (accessed January 20, 2025, 2025).