Journal of Electronic Research and Application, 2026, Volume 10, Issue 1
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Design and Implementation of a Vertical Search Engine for the Fisheries Domain

Zhiqiang Zhang*

School of Artificial Intelligence, Zhejiang Dongfang Polytechnic, Wenzhou 325000, Zhejiang, China

**Corresponding author: Zhiqiang Zhang, cqcjzd123@126.com*

Abstract: This paper aims to address the issues of insufficient professionalism and precision in information retrieval within the fisheries domain by designing and implementing a vertical search engine specifically for fisheries. The system employs a specialized structure, integrating a fisheries terminology-based word segmentation algorithm, a query expansion mechanism based on knowledge graphs, and a multi-source fisheries data collection and processing workflow. It establishes a domain knowledge framework that includes various entities such as fishing gear, fish species, fishing grounds, and legal regulations. The system enhances text representation and retrieval relevance by using word segmentation techniques that combine mutual information and left-right entropy, as well as TF-IDF weighting and the vector space model. Experiments show that the system's response time for retrieving information from fisheries-specific databases is within 1 second, which is a significant improvement compared to traditional search engines. The system demonstrates good domain adaptability and practical value.

Keywords: Fisheries domain; Information retrieval; Vertical search engine; Word segmentation algorithm; Knowledge graph

Online publication: February 27, 2026

1. Introduction

The development of information technology and the growth of internet data have made the efficient and accurate acquisition of knowledge in a specific field the core of industry intelligence and digital transformation. The information ecosystem of the fisheries sector is becoming increasingly complex and diverse. Currently, industry professionals rely heavily on general-purpose search engines to obtain specialized knowledge [1]. However, these search engines have significant shortcomings in the fisheries domain. On one hand, their precision is low, as the search results are often cluttered with irrelevant or low-quality commercial information, making it difficult to find authoritative and accurate content. On the other hand, the organization of information is lacking, as the results are not displayed according to the logical classification within the fisheries domain. This hinders the speed

of information acquisition and negatively impacts production decision-making, technology dissemination, and industry management [2].

This paper proposes a fisheries-specific word segmentation method that combines a domain dictionary with statistical models, effectively correcting improper segmentation of compound professional terms such as “fish-finding sonar”. It also establishes a fisheries knowledge graph and a parametric query analyzer that can accurately parse and expand structured query statements like “trawl mesh size ≥ 50 meters”. Furthermore, a specialized search framework has been developed to integrate fisheries information from various sources and perform real-time incremental indexing, enhancing the system’s search adaptability for multiple application scenarios, including equipment parameters, fishery conditions, and legal regulations, while maintaining search speed.

2. System architecture

2.1. System design

The structure of the collection system for the vertical search engine in the fisheries domain is shown in **Figure 1**. First, enter the system management interface to configure the parameters of the crawling task for the fisheries domain websites. The specific configuration process is omitted. The crawling process is controlled by the collection scheduling controller: when the collection scheduling controller is initialized, it loads the configured collection parameters, and at the same time, initializes the fisheries domain URL database, related logs, crawling scope restrictions, URL scheduler, processing pipeline, and thread pool used during the collection period. After starting the thread pool and the URL scheduler, the data collection task for the fisheries domain begins [3].

The URL scheduler provides the collection threads with links to be crawled, which are divided into two types: links in the to-be-crawled queue and links that have been crawled. The system uses an embedded database to store the links of fisheries-related web pages to be crawled and those that have been crawled separately. The collection threads obtain the content of fisheries web pages based on the provided links.

The crawling task is collaboratively completed by five processing pipelines: the pre-crawling pipeline is responsible for judging the crawling conditions and will only obtain web page content if it meets the specific crawling conditions of the fisheries domain, for example, it can adjust the priority for authoritative fisheries websites; the protocol processing pipeline is in charge of parsing various network transfer protocols; the content extraction pipeline is responsible for extracting the core content of fisheries-related web pages; the storage writing pipeline writes the extracted fisheries data into the storage system; and the link analysis pipeline identifies new fisheries-related links from the web page content and adds them to the URL scheduler for continuous crawling [4].

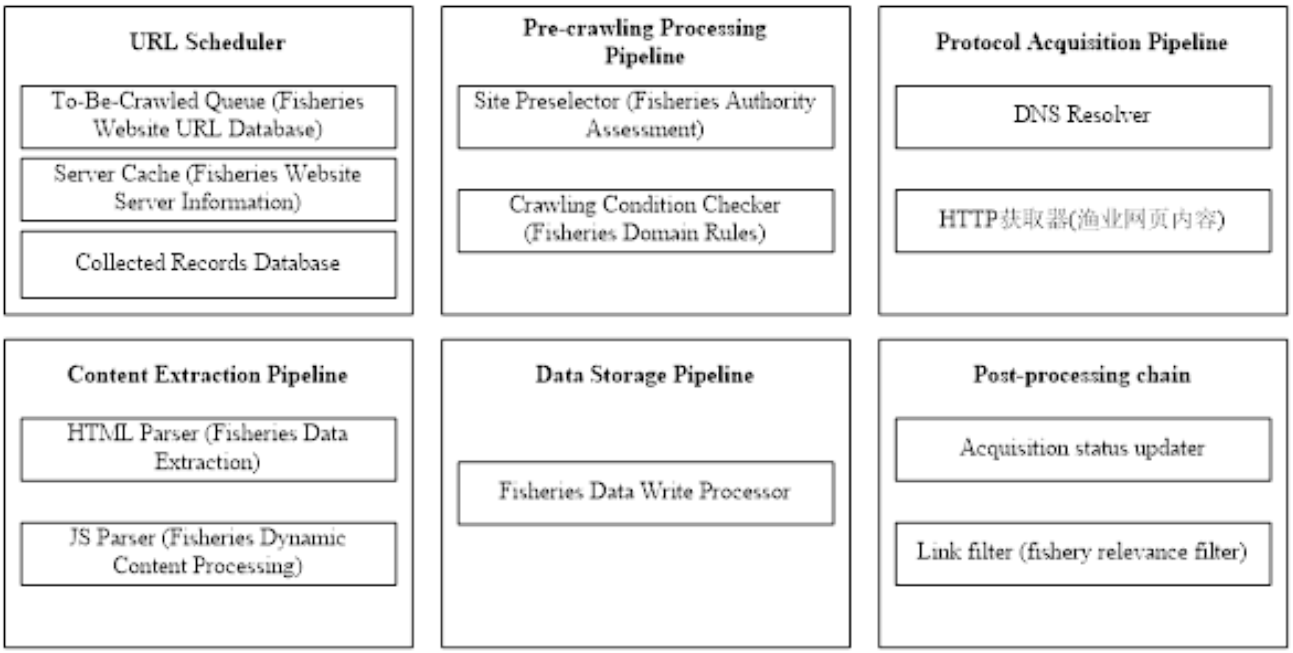

Figure 1. The structure of the collection system for the vertical search engine in the fisheries domain.

2.2. Main functional design

The specialized search engine for the fisheries industry adopts a professional system architecture design, and the establishment of a fisheries knowledge graph is a key part of this system. This section builds a knowledge graph based on domain ontology, using domain ontology modeling tools to determine entities such as fishing gear, fish species, fishing vessels, fishing grounds, and related laws and regulations, as well as their interrelationships. Moreover, an entity recognition engine that combines deep learning technology and rule-based methods is employed to accurately extract fisheries-specific terminology. Subsequently, a relationship discovery system is used to parse the semantic connections between entities, ultimately forming a structured knowledge system in the fisheries domain.

The system is equipped with an intelligent query understanding and processing module, which has been specifically optimized for the characteristics of fisheries queries. This module can identify the user's query intent, such as queries for fishing gear parameters, searches for fishery information, or searches for policies and regulations. It also utilizes a fisheries thesaurus and knowledge graph to expand the scope of the query. The parametric query parser is particularly suitable for parsing query statements containing technical conditions, such as "trawl mesh size ≥ 50 meters". The document processing and indexing section is responsible for standardizing and organizing various sources of fisheries data. It can interpret professional documents such as fishing vessel design drawings, catch statistics tables, and scientific research reports. Using fisheries spatiotemporal data extraction tools, it accurately obtains core data such as the location coordinates and operation times of fishing grounds, thereby establishing an inverted index system that adapts to the characteristics of industry terminology [5]. Subsequently, the precise search and ranking mechanism, while considering the professionalism of terms, the relevance between texts, and the novelty of information, categorizes based on the credibility level of fisheries information sources and provides personalized ranking feedback to different user groups.

The document processing and indexing section is responsible for standardizing and organizing various sources of fisheries data. It can interpret professional documents such as fishing vessel design drawings, catch

statistics tables, and scientific research reports. Using fisheries spatiotemporal data extraction tools, it accurately obtains core data such as the location coordinates and operation times of fishing grounds, thereby establishing an inverted index system that adapts to the characteristics of industry terminology. Subsequently, the precise search and ranking mechanism, while considering the professionalism of terms, the relevance between texts, and the novelty of information, categorizes based on the credibility level of fisheries information sources and provides personalized ranking feedback to different user groups.

In terms of system integration, various data access interfaces have been designed to connect with domestic and international fisheries databases, Internet of Things monitoring data, and satellite remote sensing data. The system also integrates third-party location-based services, marine climate predictions, and market intelligence on aquatic products. Using standard interfaces, it enables data and service interaction with fisheries management systems, scientific research platforms, and corporate applications.

2.3. Design of the fisheries domain word segmentation plugin

Word segmentation is crucial for a vertical search engine specialized in the fisheries domain, as the method of segmentation directly affects the effectiveness of information retrieval related to fisheries. The word segmentation mechanism of the system described in this paper mainly encounters the following three technical challenges.

The built-in word segment of ElasticSearch has poor adaptability to Chinese professional terms in fisheries, often breaking compound words into individual characters. For example, the term “fish-finding sonar” is incorrectly divided into “probe”, “fish”, “sound”, and “sodium”.

Chinese word segmentation is inherently more complex than English segmentation, which can be done simply based on spaces or punctuation marks. In contrast, Chinese segmentation must utilize semantic and domain knowledge, especially in professional fisheries literature, where the difficulty is greater.

During the retrieval of fisheries catch equipment, when users query based on parameters, inaccurate segmentation often leads to matching irrelevant fisheries data, thereby affecting the precision of the search results.

In light of the above reasons, the system adopts a word segmentation method that combines mutual information and left-right entropy. Mutual information is used to evaluate the appropriateness of character combinations forming professional fisheries terms, while left-right entropy is employed to analyze the diversity of surrounding characters and words. A higher entropy value indicates that the term frequently appears in fisheries literature and is more likely to be an effective professional fisheries term.

In the vertical search engine for the fisheries industry, even after text segmentation, computers still cannot understand fisheries-related natural language as humans do. This requires transforming the text into a feature representation form [6]. Using the Vector Space Model (VSM), fisheries-related texts can be converted into n -dimensional feature vectors, where each dimension of the vector represents a feature in the text, and each feature has a corresponding weight, that is:

$$D = D\left(t_1 w_1; t_2 w_2; \cdots; t_n w_n\right) \tag{1}$$

In Equation (1): The feature item t_n is the basic unit of fisheries text information, referring to words here, such as “trawl”, “fishing season” and “sonar” etc.; while the weight w_n is the measure of the importance of the word in the fisheries text D.

TF-IDF is a weighting technique used to evaluate the value of fisheries literature, aiming to quantify the importance of terms in professional documents within the inverted index. The importance of a word t_i in a specific fisheries document can be expressed by the following formula:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (2)$$

In Equation (2), $n_{i,j}$ represents the frequency of word d_j appearing in the fisheries document d_j, while $\sum_k n_{k,j}$ is the total number of occurrences of all words in document d_j. The calculation method is as follows:

$$idf_i = \log \frac{|D|}{u_i + 1} \quad (3)$$

In Equation (3): $|D|$ represents the total number of documents in the fisheries domain; u_i represents the number of documents containing the term t_i. By multiplying the results of Equation (2) and Equation (3), the value of $TF - IDF_{i,j}$ is obtained. This value can be used to extract keywords from the inverted index of fisheries documents:

$$TF - IDF_{i,j} = tf_{i,j} \times idf_i \quad (4)$$

This method is conducive to identifying the uniqueness of fisheries-specific terms such as "encircling net fishing", "catch statistics", and "fishing vessel navigation" in the literature, thereby improving the search effectiveness of the vertical search engine.

Categorical search, as one of the core functions of the vertical search engine for the fisheries domain, is systematically and logically designed. **Figure 2** illustrates the process, reflecting the systematic and orderly nature of searching for fisheries-specific knowledge. Users first arrive at the categorical search page and select the primary fisheries category they wish to query. Based on the selected primary category, the system loads the relevant secondary classification directories and subsequently displays the fisheries professional database under that category. Users can decide whether to include subcategories, and the system will display the corresponding content according to the user's selection.

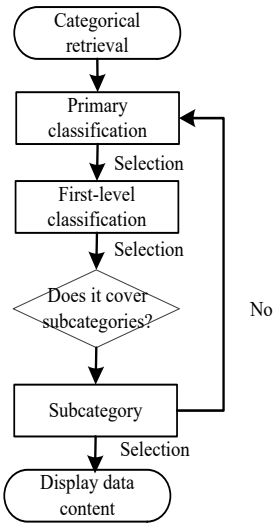

Figure 2. Process design.

At the database design level, the system employs MySQL for structured data storage, beginning with the design of an E-R diagram. The E-R model is shown in **Figure 3**, depicting the relationships between various fisheries entities within the system. The system encompasses equipment data for multiple fishing operation

methods, which are associated with specific categories such as deck machinery, fishing gear, and buoys. Each category has detailed attributes, including specific parameters and equipment codes. The entity relationships in the system reflect data structure associations in multiple aspects, such as user information, equipment data, and system management, ensuring the accuracy and completeness of fisheries data queries.

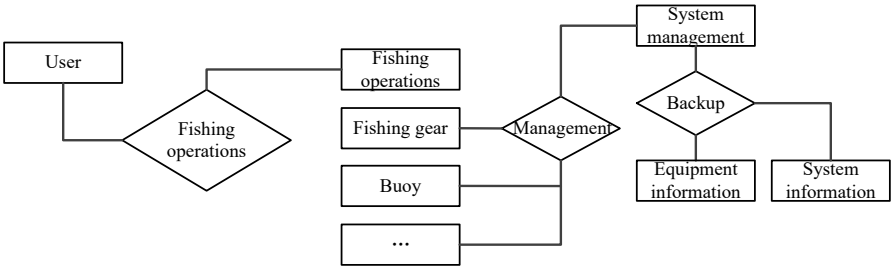

Figure 3. E-R model.

3. Experimental results analysis

Based on the technical solution proposed in this study, a prototype vertical search engine for the fisheries domain was developed. The experiments were conducted on a software platform comprising JDK 1.8, Tomcat 8.5, and Spring Boot 2.x, as well as a hardware environment equipped with an Intel Core i7-10700 processor (2.9GHz) and 16GB of memory. The system was tested on a dataset containing approximately 500,000 fisheries-related documents, including research papers, technical manuals, industry reports, policy regulations, and equipment parameters. A specialized index was built for this dataset, and corresponding retrieval services were provided. **Table 1** shows the differences in key modules of the search engine. The system has made significant improvements and enhancements in data source specificity, index structure optimization, fisheries domain word segmentation algorithms, and specialized query analysis compared to traditional search engines.

Table 1. Differences in key modules of the search engine

Key technologies	The fisheries vertical search engine designed in this paper	Traditional general-purpose search engines
Data sources	Supports multi-source data interfaces, enabling targeted collection from fisheries-specific databases, industry websites, and scientific literature	Mainly rely on general web crawlers, which cover a wide range but lack domain specificity
Indexing content	Intelligently filters and cleans fisheries data, indexing only domain-related structured and semi-structured information	Typically index all publicly available information on the web, lacking domain filtering mechanisms
Index structure	Adopts an inverted index optimized for fisheries terminology, supporting real-time incremental updates and historical version management	Mostly use traditional inverted indexes, with longer update cycles and no support for fine-grained incremental indexing
Word segmentation algorithms	Employs a Chinese word segmentation algorithm that integrates a fisheries domain dictionary with statistical models, supporting professional term recognition and new word discovery	Mainly depend on general word segmentation models, which have lower accuracy in recognizing fisheries-specific terms and compound words
Query analysis	Includes built-in interfaces for fisheries query semantic understanding and expansion, supporting parametric search, multimodal querying, and intent recognition	Usually based on keyword matching, with a single query method and lacking domain semantic understanding and expansion capabilities

To evaluate the performance of the fisheries vertical search engine, the information collection time and number of files for multiple typical fisheries data sources were statistically analyzed. The collection time and data scale of fisheries data sources are shown in **Table 2**. The data collection time is basically positively correlated with the data scale. For government information and scientific research databases, the system can complete the collection in a relatively short time, with the shortest being 137.438 seconds. However, for large-scale data sources such as the Global Fisheries Remote Sensing System, the collection time is correspondingly longer due to the large data volume and complex structure. The system has the capability to handle large-scale fisheries data, but there is still room for improvement in the collection efficiency of ultra-large-scale data sources. In the future, the system can be further improved through distributed collection and incremental update strategies.

Table 2. Collection time and number of files for fisheries data sources

Fisheries data sources	Information collection time / s	Number of files and documents / pieces
China fisheries government website	137.438	7339
Aquatic science database	235.140	30702
Distant water fisheries Information platform	510.375	52753
Fisheries equipment Technology library	1633.245	170057
Global fisheries remote Sensing monitoring system	3415.167	38 510

The fisheries keyword search time and result quantity are shown in **Table 3**. The test selected common high-frequency search keywords in the fisheries domain for querying. All search response times were within 1 second. For example, the search for “trawl” took 0.047 seconds and yielded 1,235 results, indicating that the system can quickly respond to users’ professional search needs. In terms of search time, the system-maintained millisecond-level responses for different keywords, demonstrating the effectiveness of the index structure. The search for “fishing vessel main engine” took relatively longer, possibly due to the involvement of numerous technical parameters and subcategories, which required more semantic expansion and relevance calculations during the search process. The number of results reflects the prevalence of the terms within the domain. “Trawl” and “fishing moratorium” are core operational methods and policy keywords, covering the most documents, which is consistent with the actual situation in the fisheries domain. Overall, the search engine meets the real-time query needs of fisheries professionals in terms of both search speed and result coverage.

Table 3. Fisheries keyword search time and result quantity

Search keywords	Search time / s	Number of search results / Entries
Trawl	0.047	1235
Tuna	0.078	892
Fishing vessel main engine	0.281	456
Sonar fish detection	0.078	678
Fishing moratorium	0.041	1024

4. Conclusion

This paper constructs a vertical search engine for the fisheries domain, realizing a full-process solution from data collection, specialized word segmentation, knowledge construction to efficient retrieval. The experiment was conducted on a dataset covering approximately 500,000 fisheries documents on the platform of JDK 1.8, Tomcat 8.5, and Spring Boot 2.x for performance verification. The test data shows that the search response time for typical fisheries keywords such as “trawl” and “tuna” is less than 0.2 s, among which the search for “trawl” only takes 0.047 s and returns 1,235 results, reflecting good real-time performance and recall ability. Comparative analysis indicates that the system outperforms traditional search engines in terms of data source specificity, index structure optimization, accuracy of domain word segmentation, and query semantic understanding. Although there is still room for improvement in the collection efficiency of ultra-large-scale data, the system has good domain applicability and retrieval performance, and can provide effective information support for fisheries research, production, and management, with promotion value and application prospects.

Funding

Scientific Research Fund of Zhejiang Provincial Education Department (2025 General Research Project): Design and Implementation of a Vertical Search Engine for the Fisheries Domain (Project No.: Y202559387)

Disclosure statement

The author declares no conflict of interest.

References

[1] Hungevu R, Lawal A, Yinusa S, et al., 2025, The Impact of Low-Cost Technological Innovations on Sustainable Fisheries for Economic Development in Developing Countries. World Journal of Advanced Research and Reviews, 25(2): 1170–1184.

[2] Sun C, Li X, Zou W, et al., 2025, Unequal Opportunities and Green Transition: A Study on the Mechanism of Green Total Factor Productivity Differences in China’s Regional Marine Economy. Regional Studies in Marine Science: 104249.

[3] Liang Y, Zhu Y, Sun Z, et al., 2023, Feasibility Assessment of a CO_2-Based Power, Cooling, and Heating System Driven by Exhaust Gas from Ocean-Going Fishing Vessel. Journal of Cleaner Production, 406: 137058.

[4] Priharanto Y, Yaqin R, Marjianto G, et al., 2023, Risk Assessment of the Fishing Vessel Main Engine by Fuzzy-FMEA Approach. Journal of Failure Analysis and Prevention, 23(2): 822–836.

[5] Koričan M, Vladimir N, Fan A, 2023, Investigation of the Energy Efficiency of Fishing Vessels: Case Study of the Fishing Fleet in the Adriatic Sea. Ocean Engineering, 286: 115734.

[6] Sathish T, Ağbulut Ü, George S, et al., 2023, Waste to Fuel: Synergetic Effect of Hybrid Nanoparticle Usage for the Improvement of CI Engine Characteristics Fuelled with Waste Fish Oils. Energy, 275: 127397.