Ontology-Driven Multi-Source Heterogeneous Data Integration Using SSN-SAO Framework for Patent Similarity Analysis

Huiqin Du, Lian Ma*
School of Information, Sichuan Vocational College of Finance and Economics, Chengdu 610101, China
E-mail: 15928798109@163.com, m18280435171@163.com
*Corresponding author

Keywords: SSN, SAO, multi source heterogeneous data, integrated technology, patent

Received: June 23, 2025

Differences in data models and standards make it difficult to directly compare, exchange, or share multi-source, heterogeneous data. Queries between data sources are also challenging, resulting in low data utilization. Therefore, this study proposes a multi-source heterogeneous data integration technology based on the concept of ontology. First, the multi-source heterogeneous data semantic integration technology based on semantic sensor networks is analyzed. Then, based on this, the subject-verb-object semantic structure is introduced and analyzed. The patent representation target is identified by combining this structure with the text characteristics of patent literature. Finally, analyze the results of the proposed technology. The results showed that the recall rate of the integrated technology, which combined semantic sensor networks and subject-verb-object semantic structures, ranged from 64.8% to 68.1%. Its F value was higher than that of the comparison technology, followed by AGDISTIS, and TagMe2 had the lowest. As the number of candidate individuals increased, the precision rate gradually rose, reaching up to 82.6% at its peak. When applied to patent processing, threshold combinations 1, 2, 5, and 9 performed better. Among them, the proportion of patent similarity repetition values in threshold combination 9 was 18%, and the proportion of patents with a similarity of 0 was 46%. After checking the patent content, it was found that its measurement results were the most accurate. Consequently, the proposed technique not only delivers state-of-the-art performance but also markedly elevates the exploitation of multi-source heterogeneous data, furnishing a robust technological backbone for both scholarly inquiry and practical deployment across relevant domains.

Povzetek: Študija predlaga ontološko zasnovano integracijo večizvornih heterogenih podatkov, ki z združitvijo semantičnih senzorskih omrežij in subject-glagol-objekt structure *preseže referenčne pristope.*

1 Introduction

With the rapid development and widespread application of information technology, human society is accumulating massive amounts of data at an unprecedented speed [1]. Data is heterogeneous due to its origin from different systems or domains. Due to the widespread deployment of devices in various parts of cities for information collection, data sources are numerous and widely distributed, resulting in multi-source data [2]. Multi Source Heterogeneous Data (MSHD) is widely present in various industries and fields, such as smart cities, healthcare, financial services, industrial manufacturing, etc., providing rich resources for data analysis and decision support [3-4]. The existence of MSHD directly leads to low system query efficiency. Therefore, effectively integrating and managing MSHD has become an urgent problem that must be solved. Many experts and scholars at home and abroad have conducted in-depth explorations in the field of MSHD. Scholars Huang and Wu proposed a two-stage adaptive ensemble method with MSHD to adaptively integrate data from different sources and distributions, and verified through empirical analysis that the proposed method could significantly improve prediction performance [5]. Thirumahal et al. proposed a machine learning based MSHD automatic integration method to address the issues of dispersed and difficult to obtain biomedical data, which could be effectively deployed and used in the healthcare field [6].

The term 'ontology' originally originated from the philosophy, but in computer science and information technology, it is used as a shared vocabulary. Among them, the Semantic Sensor Network (SSN) ontology is a tool specifically designed to describe elements such as sensors, related processes, and research interests [7]. The Subject Action Object (SAO) ontology is a semantic description structure formed based on the three types of subject verb object structures. As an ontology specifically designed to describe sensor observation data, it provides a powerful tool for semantic integration of sensor data [8]. In recent years, many scholars have conducted in-depth explorations on SSN ontology and SAO ontology. Chen et al. proposed a threshold pruning algorithm with SSN, which could improve retrieval efficiency and response time [9]. The Palmblad team applied SAO to chemical experimental literature and proposed an integrated method with multi-source data resources, which could be directly applied to chemical experimental literature retrieval [10].

In summary, with the theoretical foundation of previous studies, this study introduces the concept of ontology and proposes an MSHD integration technology with SSN and SAO (SSN-SAO-MSHD), aiming to achieve automatic data integration and association by adding semantic layers to MSHD. This study innovates by using the SSN ontology to describe sensor network structures and observed data attributes, as well as the SAO ontology to semantically annotate the observed data. This approach converts sensor data from various sources and formats into a unified semantic model, thereby improving the efficiency of MSHD recognition. The summary table of relevant literature is shown in Table 1.

Table 1: Summary of relevant literature

Literature	Method	Application field	Core technology	Performance results
[5]	Two-stage adaptive integration	Adaptive integration scenarios for multi-source heterogeneous data	Two-stage adaptive integration framework for multi-source heterogeneous data	Significantly improve the accuracy of prediction
[6]	Machine learning automatic integration method	Biomedical field	Machine learning algorithm-driven automatic integration technology	The deployment in the healthcare field is effective
[9]	SSN threshold pruning algorithm	Semantic sensor network data retrieval scenario	SSN ontology and threshold pruning optimization technology	Improve retrieval efficiency and response time
[10]	SAO Chemical literature integration method	The field of chemical experiment literature	SAO ontology, multi-source data resource integration framework	It is directly applicable to chemical literature retrieval
Our	A multi-source heterogeneous data integration method based on SSN and SAO	Patent field	Semantic integration technology integrating SSN and SAO	It can improve the utilization rate of multi-source heterogeneous data

In Table 1, there are still obvious gaps in current patent research. The accuracy and efficiency of semantic integration of patent data must be improved due to the dynamic evolution of technical terms and significant differences in cross-source formats. To this end, the research proposes a multi-source heterogeneous data integration technology based on SSN and SAO, aiming to achieve automatic data integration and association by adding a semantic layer to multi-source heterogeneous data. The study clarifies the following research questions. First, compared with existing methods such as AGDISTIS and TagMe2, does the multi-source, heterogeneous data integration technology that combines SSN and SAO (SSN-SAO-MSHD) improve the precision and recall rates in patent similarity tasks? Does the introduction of the SAO semantic structure reduce redundant results in patent similarity measures? In response to the above issues, the research is expected to achieve the following results. In the patent dataset, the SSN-SAO-MSHD technology has a precision rate of at least 80%, a recall rate of at least 65%, and an F-value that is more than 5% higher than that of existing mainstream methods. Through the precise analysis of the relationship by SAO, the proportion of patent similarity repetition values has been reduced to below 10%. The innovation point of this research lies in using the SSN ontology to describe the structure of the sensor network and the attributes of the observed data. Meanwhile, the SAO ontology is used to provide semantic annotations for technical features, actions, and application objects within patent texts. This converts heterogeneous patent data from multiple sources and various formats into a unified dataset. This will enhance the efficiency of identifying technical associations in patent data.

2 Methods and materials

2.1 MSHD semantic integration technology with SSN

MSHD integration technology can handle diverse data sources and complex data structures, but it lacks in data semantic integration. Therefore, this study uses SSN ontology for optimization [11]. By introducing SSN ontology, a clear and structured data representation framework can be constructed, thereby improving semantic consistency and accuracy in the data integration process. The schematic diagram of the SSN body and its longitudinal and transverse modules is shown in Figure 1.

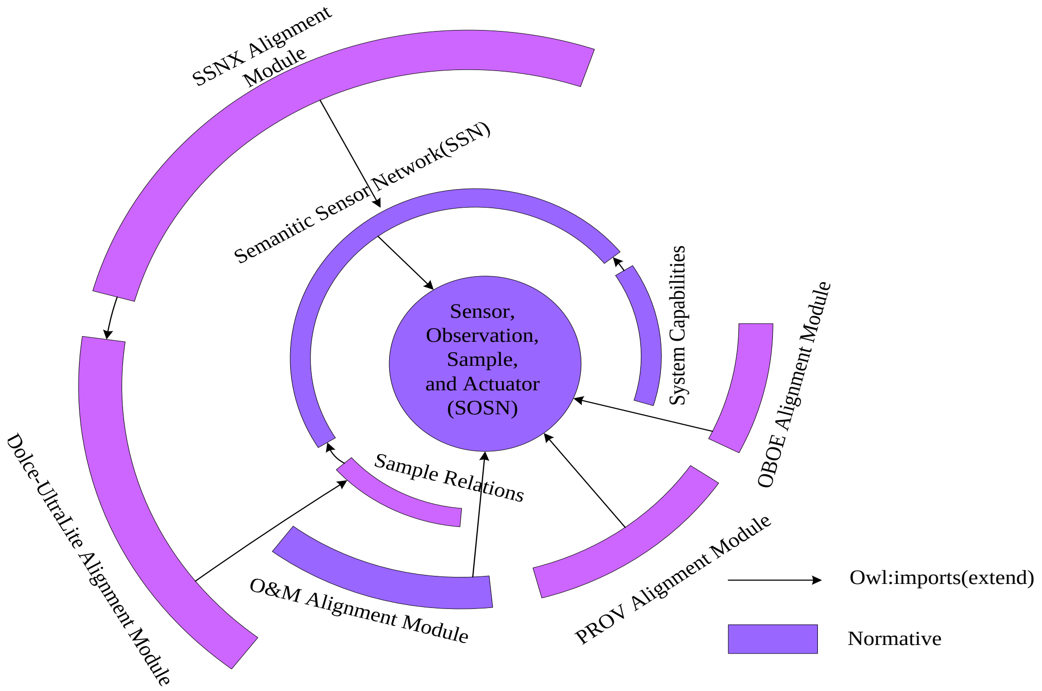


Figure 1:Diagram of the SSN body and its vertical and horizontal modules.

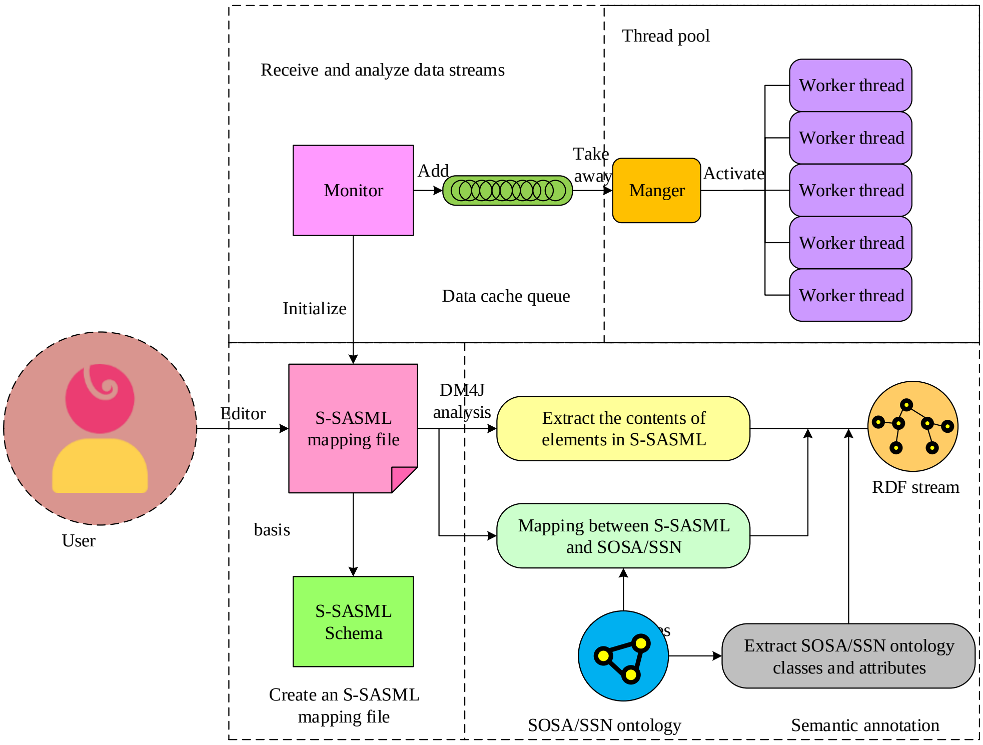


Figure 2:Semantic integration framework of MSHD with SSN.

In Figure 1, the SSN ontology is divided into modules vertically, and each module is interdependent through the owl:import statement, forming a hierarchical structure [12]. Horizontal modules are modules divided horizontally in the SSN ontology, and there is no strict hierarchical structure between each module. MSHD language, due to its wide range of sources, diverse types, and containing a large amount of spatiotemporal information and dynamic data, makes its processing and analysis more complex. Therefore, this study introduces the SSN ontology and extracts key information from MSHD. The MSHD semantic integration technology framework based on SSN is shown in Figure 2.

In Figure 2, during the data reception and analysis phase, semantic integration technology utilizes a listener mechanism to monitor and manage the data streams generated by sensors in real-time. In the face of the large amount of data streams that sensors may generate instantly, a queue mechanism is adopted to temporarily store data to prevent data loss. To improve the real-time performance of the system, thread pooling technology has been introduced to achieve high concurrency data processing. The management thread is responsible for retrieving data from the data cache queue and assigning it to idle worker threads for processing. Through the S-ASML pattern, mapping files are created to describe the semantic relationships between different data sources. XML parsing tools such as DOM4J are used to parse the mapping files, and the corresponding relationships between elements are extracted. During the semantic annotation stage, researchers use semantic web tools, such as Jena, to parse the SSN ontology and extract relevant class and attribute information. MSHD is mapped onto the SSN ontology, and the raw data is parsed, cleaned, and transformed to meet the requirements of the SSN ontology. Meanwhile, the semantic relationships between data also need to be considered to ensure the accuracy and completeness of the mapping. Figure 3 shows the correspondence between data flow, S-ASML elements, and SSN ontology.

In Figure 3, AreaID, NodeID, and DeviceID identifiers are represented by their corresponding classes and properties in SOSA/SSN. Each sensor device can be regarded as a Sensor instance, belonging to a specific Platform and corresponding to AreaID, NodeID, and DeviceID through the identifier attribute. The position, observation event, and type attributes of the sensor correspond to the position attribute, observation event type, and sensor type of the Sensor class in SSN [13]. When quantifying the degree of heterogeneity between ontologies, the concept of recall cannot be used. Therefore, this study assumes that it is necessary to quantify the sets of similar entities E_r and E_1 in two ontologies, and correspond them according to the entity matching cardinality of 1:1. The expression for the quantified degree of heterogeneity is shown in equation (1).

$$H_{syntax} = \frac{\min\{|E_r|,|E_1|\} - |E_{mapped}|}{\min\{|E_r|,|E_1|\}} \qquad (1)$$

In equation (1), $|E_r|$ and $|E_1|$ are respectively the cardinality of E_r and E_1). $|E_{mapped}|$ represents the cardinality of the same entity in two sets of entities. If the heterogeneous data is different, then the choices of sets E_r and E_1 are different. After determining the heterogeneous values between ontologies, this study maps the heterogeneous values to the integrated weights of different similarity measurement methods to establish a regression model of "ontology heterogeneous values integrated weights" [14]. The model expression is shown in equation (2).

$$y = f(x,e,f,v) = e \times f^{x} + v \qquad (2)$$

For exponential function fitting, the least squares method is used to minimize the sum of squared errors to find the optimal function match for the data. Assuming the observed values are (x, y), the expression of the function is shown in equation (3).

$$y = f(x,w) \qquad (3)$$

In equation (3), w represents the undetermined parameter. To find the optimal estimate of the parameter w of the function y , this study assumes that the observation data of group o is (x_u, y_u), and solves the objective function expression as shown in equation (4).

$$Z(y, f(x,w)) = \sum_{u=1}^{o} [y_u - f(x_u, w_u)]^2 \qquad (4)$$

This study uses Euclidean distance to obtain the least squares function, and the expression of the least squares method is shown in equation (5).

$$\min f(x) = \sum_{u=1}^{o} Z_l^2(x) = \sum_{u=1}^{o} Z_l^2 [y_u - f(x_u, w_u)] = \sum_{u=1}^{o} [y_u - f(x_u, w_u)]^2 \quad (5)$$

In equation (5), $Z_l(x)$ represents the residual function.

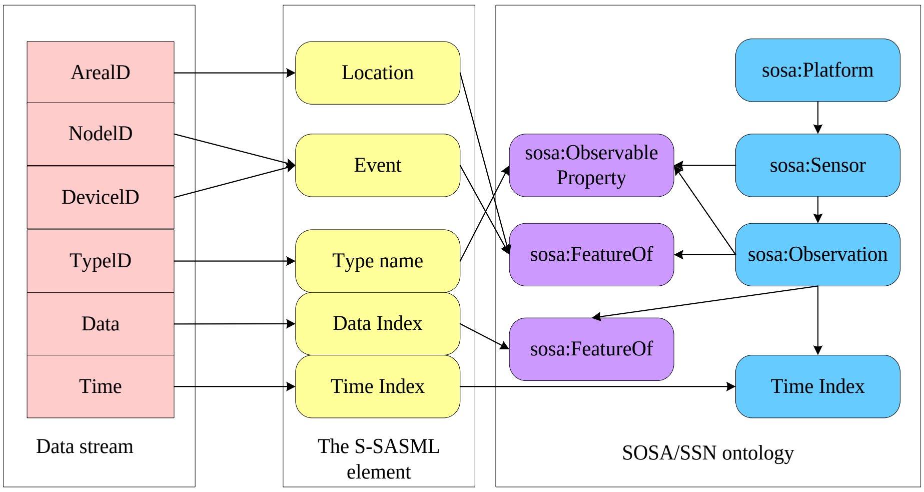


Figure 3:Mapping relationship between data flow, S-SASML element and SSN ontology.

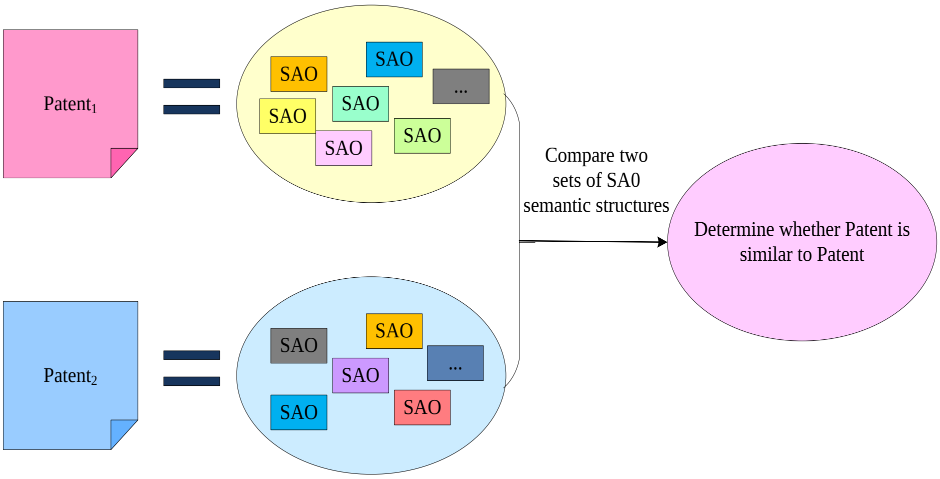

Patent$_1$
SAO
SAO
...
SAO
SAO
SAO
SAO
Patent$_2$
SAO
SAO
SAO
...
SAO
SAO
Compare two sets of SA0 semantic structures
Determine whether Patent is similar to Patent

Figure 4:SAO semantic structure diagram.

2.2 Application of SSN-SAO-MSHD

Since different sensors and data sources may adopt their own independent data models and data standards, it is impossible to directly compare, exchange and share data. Therefore, this study introduces the SAO structure in the MSHD integration technology of SSN to provide a unified data model for annotated sensor data. This allows computers to understand and automatically integrate the sensor data, solving the problem of inconsistent data models. SAO semantic structure is an important model in natural language processing and knowledge representation, which can simplify complex sentences or text fragments into a structured form composed of three basic components. The schematic diagram of SAO semantic structure is shown in Figure 4.

In Figure 4, the S structure and O structure are usually composed of several words, collectively describing the subject and object of the action. The A structure is usually represented by a verb or verb phrase, describing the specific action performed by the subject on the object. In the S structure, assuming two concepts are S_T and S_D, the matching value calculation equation for the SAO structure is shown in equation (6).

$$Sim(S_T, S_D) = \frac{2IC(common(S_T, S_D))}{IC(S_T) + IC(S_D)} \quad (6)$$

In equation (6), $common(S_T, S_D)$ represents the commonality between two concepts. IC represents the function that calculates the information content of the concepts. Assuming the concept S_C, the expression of its IC function is shown in equation (7).

$$IC(S_C) = -\log \frac{|leaves(S_C)| / |subsumers(S_C)| + 1}{\max\_leaves + 1} \quad (7)$$

In equation (7), $leaves(S_C)$ represents the total number of leaf concepts under concept S_C. $|subsumers(S_C)|$ represents the total number of parent concepts of concept S_C and S_C. $\max\_leaves$ is the total number of leaf concepts in the technical point ontology. In practical applications, each piece of information contains multiple technical points, so the corresponding set of technical points needs to be constructed. Assuming that the technical point for demand information is S_T and the technical point for supply information is S_D, the Cartesian product method is used to calculate the similarity between each technical point, and a matrix M_1 of m×n is constructed. The expression of M_1 is shown in equation (8).

$$M_1 = \begin{bmatrix} Sim(S_{T\_1}, S_{D\_1}) & Sim(S_{T\_1}, S_{D\_2}) & \cdots & Sim(S_{T\_m}, S_{D\_n}) \\ Sim(S_{T\_2}, S_{D\_1}) & Sim(S_{T\_2}, S_{D\_2}) & \cdots & Sim(S_{T\_m}, S_{D\_n}) \\ \vdots & \vdots & \ddots & \vdots \\ Sim(S_{T\_m}, S_{D\_1}) & Sim(S_{T\_m}, S_{D\_2}) & \cdots & Sim(S_{T\_m}, S_{D\_n}) \end{bmatrix} (8)$$

In equation (8), $Sim(S_{T\_i}, S_{D\_j})$ represents the similarity between technical point S_{T_i} in set S_T and technical point S_{D_j} in set S_D. The standardized equation for technical point G is shown in equation (9).

$$Sim(S_T, S_D) = \frac{(m+n)\sum_{k=1}^{K} Sim_k}{2mn} \quad (9)$$

At the grammatical level, it is assumed that A_T word set is a feature of technical performance. A_D word set is a feature of technical problems. Similar calculations are performed on A_T and A_D, and the calculation equation is shown in equation (10).

$$Sim(A_T, A_D) = \frac{|A_T \cap A_D|}{|A_T \cap A_D| + 0.5|A_T / A_D| + 0.5|A_D / A_T|} \quad (10)$$

In equation (10), $|\ |$ represents the set potential, and $/$ represents the difference set. When the number of shared words in two word sets increases, the syntactic similarity between the indicator word sets becomes higher, and the deep semantics of the words themselves will deviate. Therefore, this study removes the common parts from the word sets and focuses on describing the words that can reflect unique semantic features to further improve the accuracy and depth of similarity calculation [15]. Assuming that the intersection number between the statistical set $\bar{A}_T$ and the set $\bar{A}_D$ is r, a new semantic similarity matrix M_2 can be calculated using the Cartesian product method, as shown in equation (11).

$$M_2 = \begin{bmatrix} Sim(\bar{A}_{T\_1},\bar{A}_{D\_1}) & Sim(\bar{A}_{T\_1},\bar{A}_{D\_2}) & \cdots & Sim(\bar{A}_{T\_1},\bar{A}_{D\_p-r}) \\ Sim(\bar{A}_{T\_2},\bar{A}_{D\_1}) & Sim(\bar{A}_{T\_2},\bar{A}_{D\_2}) & \cdots & Sim(\bar{A}_{T\_2},\bar{A}_{D\_p-r}) \\ \vdots & \vdots & \ddots & \vdots \\ Sim(\bar{A}_{T\_p-r},\bar{A}_{D\_1}) & Sim(\bar{A}_{T\_p-r},\bar{A}_{D\_2}) & \cdots & Sim(\bar{A}_{T\_p-r},\bar{A}_{D\_p-r}) \end{bmatrix}$$

(11)

In equation (11), $Sim(\bar{A}_{T\_i},\bar{A}_{D\_j})$ represents the semantic similarity between vocabulary $\bar{A}_{T\_i}$ in $\bar{A}_T$ and vocabulary $\bar{A}_{D\_j}$ in $\bar{A}_D$. The cosine distance is used to obtain the similarity between words, assuming that a_i and b_i are word vectors of $\bar{A}_{T\_1}$ and $\bar{A}_{D\_1}$, respectively. The equation for calculating the similarity between $\bar{A}_{T\_1}$ and $\bar{A}_{D\_1}$ is shown in equation (12).

$$Sim(\bar{A}_{T\_1},\bar{A}_{D\_1}) = \frac{\sum_{i=1}^{h}(a_i \times b_i)}{\sqrt{\sum_{i=1}^{h}(a_i)^2} \times \sqrt{\sum_{i=1}^{h}(b_i)^2}} \quad (12)$$

In equation (12), h represents the dimensionality of the word vector. The weighted average method is used to standardize the similarity results between the word set $\bar{A}_T$ and $\bar{A}_D$, and the calculation equation is shown in equation (13).

$$Sim_{SET}(A_T, A_D) = Sim_{SET}(\bar{A}_T, \bar{A}_D) = \frac{(p+q)(r+\sum_{m=1}^{M} Sim_m)}{2pq} \quad (13)$$

This study organically integrates syntactic similarity and semantic similarity information to calculate the semantic similarity between technical performance and technical problems in supply and demand information. The calculation equation is shown in equation (14).

$$Sim(A_T, A_D) = 0.5Sim_{RE}(A_T, A_D) + 0.5Sim_{SET}(A_T, A_D)$$

(14)

Regarding technology supply information C_T and demand information C_D, this study also uses the Cartesian product method to calculate and obtain the similarity matrix M_3 of technology supply and demand information, as shown in equation (15).

$$M_3 = \begin{bmatrix} Sim(C_{T\_1},C_{D\_1}) & Sim(C_{T\_1},C_{D\_2}) & \cdots & Sim(C_{T\_1},C_{D\_f}) \\ Sim(C_{T\_2},C_{D\_1}) & Sim(C_{T\_2},C_{D\_2}) & \cdots & Sim(C_{T\_2},C_{D\_f}) \\ \vdots & \vdots & \ddots & \vdots \\ Sim(C_{T\_T-q},C_{D\_1}) & Sim(C_{T\_q},C_{D\_2}) & \cdots & Sim(C_{T\_q},C_{D\_f}) \end{bmatrix}$$

(15)

As patent data also belongs to MSHD, this study applies the SSN-SAO-MSHD to the patent field, providing strong support for patent analysis. Due to the static nature of the text and the non-real-time characteristics of the patent data stream, the study adapts the SSN framework to align with the patent data. This involves mapping the technical disclosure text in the patent data to the observation data in the SSN and corresponding the patent applicant/patentee to the sensor node of the SSN. The patent publication/application time is converted to a temporal attribute of SSN, the classification of the technical field to which the patent belongs is mapped to AreaID, and the patent document carrier is corresponded to DeviceID. Through concept transformation, the research shows that the sensor-centered SSN framework can accurately represent the multi-source attributes and semantic associations of patent data. At the same time, combined with the in-depth analysis of the patent text by the SAO ontology, a semantic integration model adapted to the patent field is constructed. The patent structure decomposition diagram of SSN-SAO-MSHD is shown in Figure 5.

In Figure 5, the technology first preprocesses the patent text, including steps such as word segmentation, part of speech tagging, and removal of stop words. Then, natural language processing tools or algorithms are used to identify SAO structures from the preprocessed patent text to determine the subject, predicate, and object in the patent. Finally, the SAO structure identified in the patent is further decomposed into three structures: S, A, and O, and the final patent content is identified and analyzed. The recognition diagram of patent characterization targets using MSHD integrated technology is shown in Figure 6.

In Figure 6, the SAO semantic structure numbered 1 appears in 5 patents. The SAO semantic structures numbered 2, 3, and 4 are distributed throughout the entire patent collection. The frequently occurring SAO semantic structure represents a general or foundational technical framework within the field, rather than a patent specific innovation point [16-17]. Number 5 with specific structure only appears in the target patent T and another patent Patent 4. From this, the SSN-SAO-MSHD can recognize the target patent, making it distinguishable from most other patents in the patent collection.

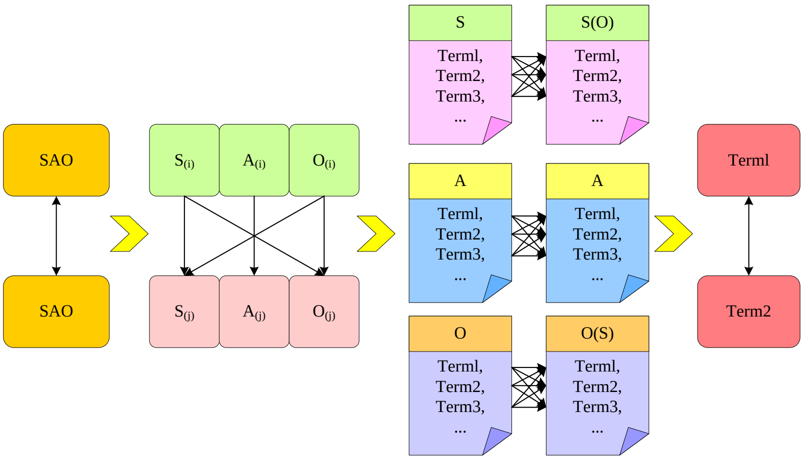


Figure 5:Patent structure decomposition diagram of SSN-SAO-MSHD.

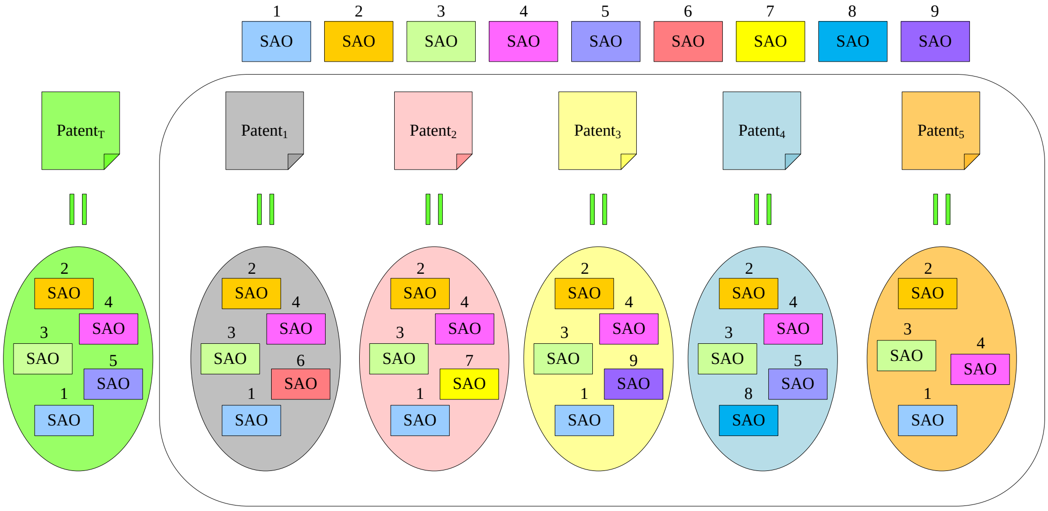


Figure 6:Identification diagram of patent characterization targets using SSN-SAO-MSHD.

3 Results

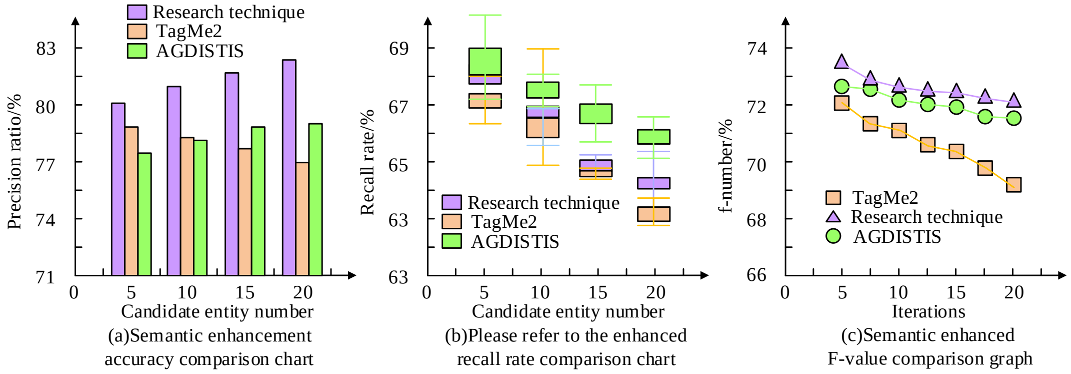
The research verifies the effectiveness of the multi-source heterogeneous data integration technology integrating SSN and SAO through hierarchical evaluation. First, the semantic integration accuracy is verified from the perspective of technical universality using indicators such as the precision and recall rates. The parallel processing ability and scalability of the system are tested through sensor nodes to support the technology's universality. Second, in response to the specific demands of the patent field, indicators such as the proportion of similarity repetition values are introduced to evaluate the practical application efficiency of the technology in the patent similarity comparison scenario. The two types of experiments form logical associations through data scale mapping and resource allocation migration. The growth in the number of sensor nodes simulates the scale expansion of multi-source patent data. The optimization results of the thread pool and cluster environment directly inform the scheduling of resources for the batch processing of patent texts. A complete verification system for basic capabilities and domain adaptation has been constructed.

3.1 Performance analysis of SSN-SAO-MSHD

To achieve the SSN-SAO-MSHD, this study conducted experiments using three types of servers: general-purpose computing servers, GPU servers, and storage servers. Table 2 shows the experimental environment settings.

Table 2: Detailed parameter table of SSN-SAO-MSHD

Server	Processor	Main frequency/Rui frequency	Nuclear number	Internal memory	GPU	Store	Bandwidth
Universal server	Intel Xeon 8269CY	2.7/3.5	24	96G	/	128G	10Gbps
GPU server	Intel Xeon 8163	2.5/3.5	24	155G	NVIDIA T4	512G	7.5Gbps
Storage server	Intel Xeon 8369B	2.5/2.7	14	32G	/	3700G	12Gbps



(a)Semantic enhancement accuracy comparison chart

(b)Please refer to the enhanced recall rate comparison chart

(c)Semantic enhanced F-value comparison graph

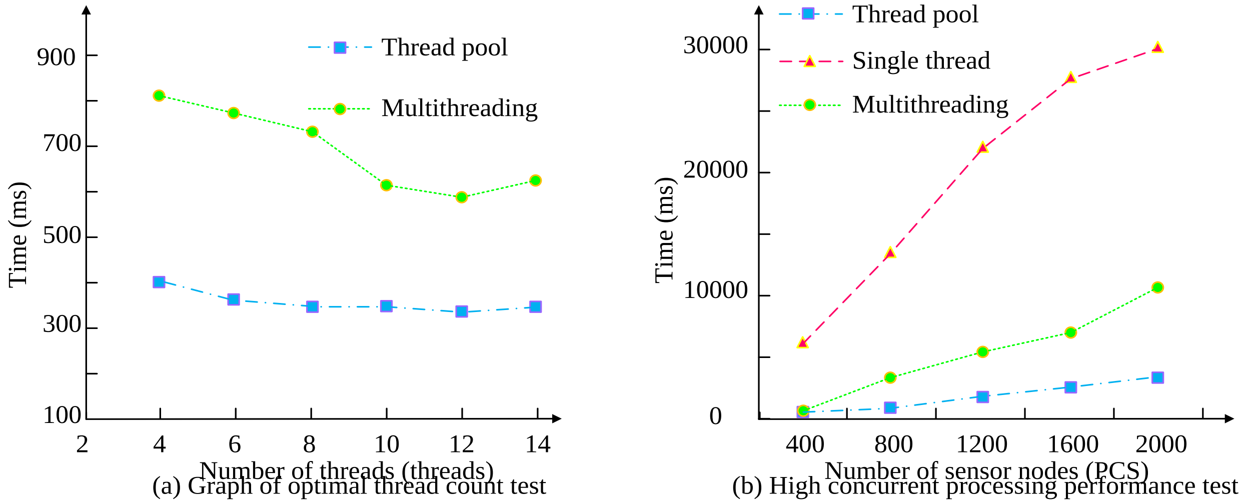
Figure 7:Performance test of SSN-SAO-MSHD.

To verify the effectiveness of SSN-SAO-MSHD, this study compared it with the well-known named entity recognition systems TagMe2 and AGDISTIS in the industry. The evaluation indicators included precision, recall, and F1 score. Figure 7 shows the performance test chart of SSN-SAO-MSHD.

In Figure 7 (a), as the candidate individuals increased, the precision of SSN-SAO-MSHD also increased. When the candidate individuals were 20, the highest precision of MSHD ensemble technology was 82.6%. In Figure 7 (b), the recall rate of SSN-SAO-MSHD was highest at 68.1% and lowest at 64.8%. In Figure 7 (c), SSN-SAO-MSHD had the highest F-value compared to the comparative technique, followed by AGDISTIS, and TagMe2 had the lowest F-value. From this, the performance of SSN-SAO-MSHD was relatively superior. To optimize and accurately test the optimal thread count configuration of the thread pool in SSN-SAO-MSHD, this study fixed the number of sensor nodes at 1200 as a benchmark condition to systematically evaluate the performance of configuring different numbers of threads in the thread pool. Then the parallel processing effects of single thread, multi thread, and thread pool were tested, as shown in Figure 8.

In Figure 8 (a), as the threads in the thread pool gradually increased, the time required for the system to process tasks initially showed a downward trend, reflecting the positive effect of enhanced parallel processing capabilities. However, when the number of threads exceeded a certain threshold, the processing time started to increase instead of decrease. In Figure 8 (b), regardless of the parallel processing strategy used, as the sensor nodes increased linearly, the system processing time showed a corresponding linear increase trend, indicating that SSN-SAO-MSHD had good scalability in high concurrency environments. To evaluate the stability and performance of SSN-SAO-MSHD, this study observed and recorded changes in processing time by adjusting the number of sensors. The experimental time was 20 seconds, and the system running time was 500 seconds. Figure 9 shows the processing time corresponding to the number of different sensors on a single machine and a cluster.

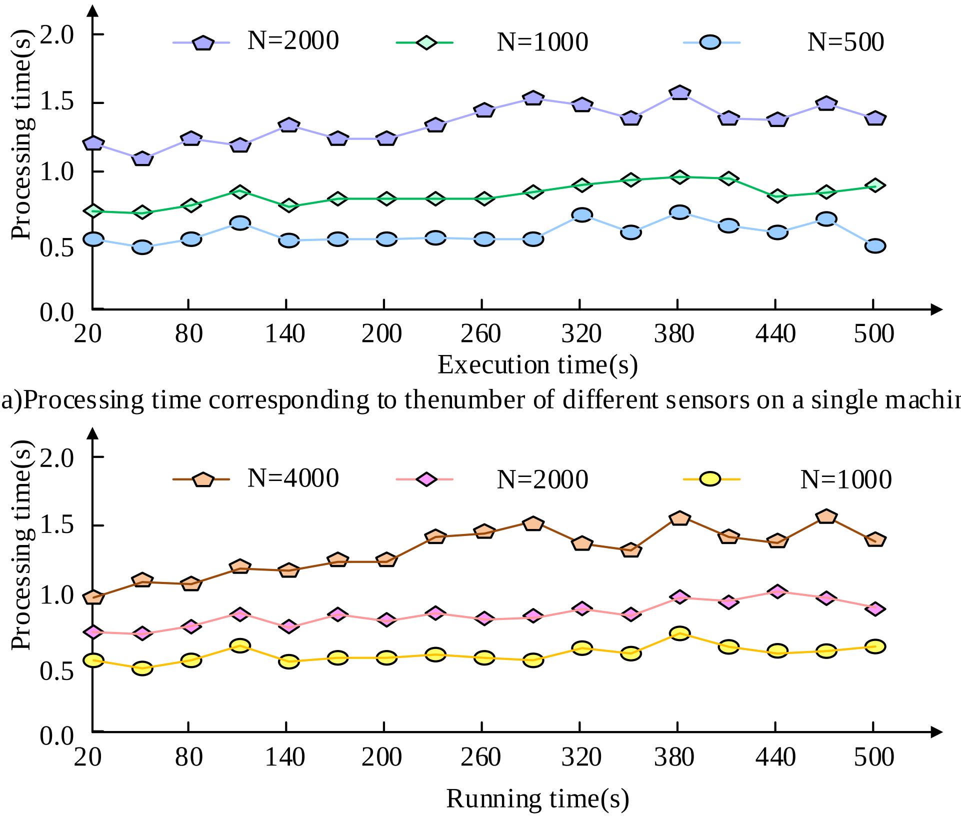
In Figure 9 (a), as the sensors increased, that was, the RDF data traffic increased, the time required for a single machine system to process these data showed a gradually increasing trend. It indicated the performance bottleneck that a single machine processing capability may encounter when facing large-scale data streams. In Figure 9 (b), within the limited 500 second operating cycle, there was a slight fluctuation in the processing time of the research technology in the cluster environment, but it did not affect the overall performance of the system. In summary, SSN-SAO-MSHD exhibited higher stability and robustness in cluster environments compared to standalone environments when processing data streams generated by a large number of sensor nodes.



(a) Graph of optimal thread count test results

(b) High concurrent processing performance test results

Figure 8:Optimal thread count test and high concurrent processing performance test.



(a)Processing time corresponding to thenumber of different sensors on a single machine

(b)Processing time of different sensor numbers in a cluster

Figure 9:Processing time diagram of the number of sensors on a single machine and a cluster.

3.2 Application effect of SSN-SAO-MSHD

To verify the application effectiveness of SSN-SAO-MSHD in patents, this study applies this technology to the processing and analysis of patent data. The selection of threshold had a significant impact on the final results during the technical implementation process. The experimental data are sourced from the Derwent Patent Intelligence Database. To ensure the optimization and accuracy of the results, 12 sets of thresholds setting schemes with different parameter combinations were designed, as shown in Table 3.

Table 3: 12 Threshold setting schemes of different parameter combinations.

Threshold combination	Word matching threshold R	SAO semantic structure matching threshold Q	Proportion of duplicate values of patent similarity	The proportion of patent similarity is 0
1	≥0.6	> 0.5	16.08%	3.15%
2	≥0.6	> 0.6	28.52%	4.96%
3	≥0.6	> 0.7	64.22%	40.26%
4	≥0.6	> 0.8	82.17%	60.24%
5	≥0.7	> 0.5	19.86%	7.95%
6	≥0.7	≥0.6	43.89%	17.19%
7	≥0.7	≥0.7	75.03%	60.13%
8	≥0.7	≥0.8	85.42%	78.26%
9	≥0.8	> 0.5	31.97%	12.59%
10	≥0.8	≥0.6	53.62%	24.82%
11	≥0.8	≥0.7	80.95%	75.02%
12	≥0.8	≥0.8	93.17%	86.70%

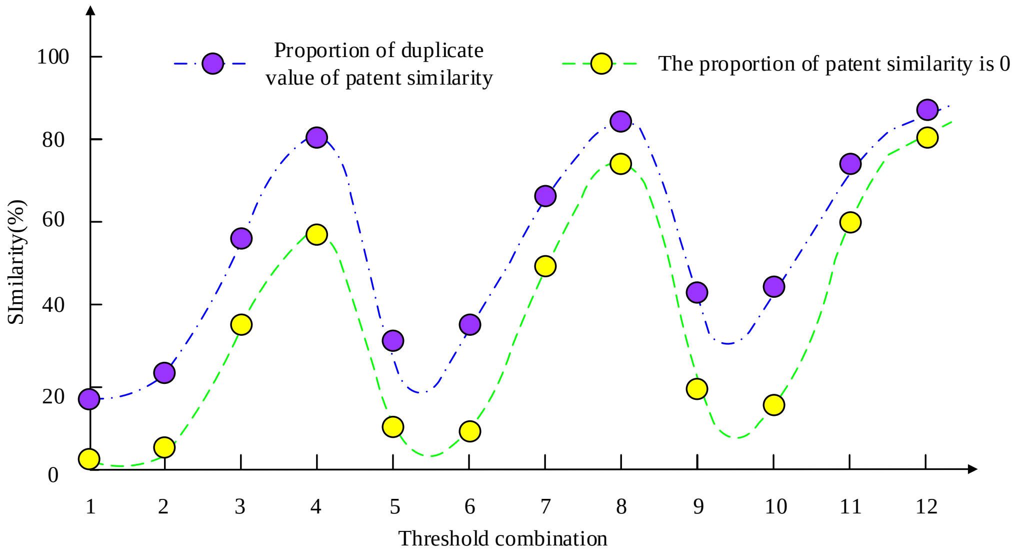


Figure 10: Patent similarity result graph corresponding to different threshold combinations.

In Table 3, under the condition of keeping the word matching threshold R constant, as the SAO semantic structure matching threshold Q in SSN-SAO-MSHD increased, the proportion of repeated similarity values significantly increased, and the proportion of patents with a similarity of 0 also showed an upward trend. To effectively distinguish the degree of similarity between different related patents and the target patent, the proportion of similarity duplicate values needed to be minimized, and the share of patents with a similarity of 0 needed to be synchronously reduced. Therefore, this study analyzed the similarity measurement results under 12 different threshold settings, as shown in Figure 10.

In Figure 10, compared to the other 8 threshold combinations, threshold combinations 1, 2, 5, and 9 had a lower proportion of duplicate patent similarity values and a proportion of patents with similarity of 0, which was in line with the optimization goal pursued by the study when selecting thresholds. Among them, the proportion of patents with similarity of 0 for threshold combination 1 was 4%, and the proportion of duplicate patent similarity values was 20%. The proportion of patents with similarity of 0 for threshold combination 2 was 6%, and the proportion of duplicate patent similarity values was 23%. 8% of patents had a similarity value of 0 for the threshold combination of 5, and 31% of patents had duplicate similarity values. The proportion of patents with similarity of 0 for threshold combination 9 was 18%, and the proportion of duplicate patent similarity values was 46%. Upon reviewing the patent content, the measurement result of threshold combination 9 was the most accurate. Therefore, when the threshold combination was 9, SSN-SAO-MSHD had a good application effect.

4 Discussion and conclusion

4.1 Discussion

Against the backdrop of rapid development of big data and the Internet of Things, the integration and processing of

MSHD is an important research direction and application field. Therefore, this study proposed an SSN-SAO-MSHD and verified its performance and practical application effects through experimental analysis.

In the performance analysis experiment, as the number of candidate individuals increased, the precision of SSN-SAO-MSHD significantly improved, reaching up to 82.6%. This level of precision is valuable for practical applications in the patent field. For core tasks such as patent infringement searches and technical similarity comparisons, a precision rate of 82.6% can significantly reduce the cost of manually screening invalid results. In addition, the integrated technology outperformed AGDISTIS and TagMe2 in terms of recall rate and F-value, with rates of 68.1% and 73.6%, respectively. The F-score of AGDISTIS was approximately 72.7%, and that of TagMe2 was about 72.3%. The research method's F-score advantage mainly stemmed from the SAO semantic structure's precise capture of the verb-object relationship in patent texts. Unlike AGDISTIS, which relied on static ontology mapping, and TagMe2, which focused on entity links, the SAO semantic structure fully explored the deep semantic associations of patent texts. This finding aligned with Yang's conclusion in "English semantic translation feature extraction," which emphasized the crucial role of deep semantic associations in enhancing task performance. Thus, it confirmed the significance of semantic structure analysis in text processing [18]. With a recall rate of 68.1%, it was estimated that about 31.9% of relevant patents were overlooked (false negatives), primarily in the area of cross-domain technology integration. The SAO semantic structure of such patents often had ambiguous expressions, leading to matching deviations. This was similar to the conclusion reached by Ben et al. in the medical monitoring system based on dynamic ontology reasoning [19]. As the threads increased, the initial processing time of the system significantly decreased, reflecting the efficiency improvement brought by parallel processing. When the number of threads exceeded a certain threshold, the processing time actually increased, mainly due to resource competition between threads and increased context switching overhead, which was consistent with the results obtained by Yu et al. [20]. This result served as a reference for system deployment and aligned with Bettaz and Maouche's viewpoint in the Internet of Things applications ontology modeling framework that resource allocation and task efficiency must be dynamically balanced. This highlighted the common principle of resource scheduling in multi-source data processing [21]. In the similarity experiment, this study optimized the accuracy of similarity measurement by adjusting the threshold Q for SAO semantic structure matching. The 12 selected threshold combinations were not randomly set but determined based on the previous sensitivity analysis. When the Q value was in the range of 0.3 to 0.8, the stability of the SAO structure showed significant fluctuations. This range covered the common distribution of the matching degree of technical terms in patent texts. Based on this, in combination with the trade-off requirements of the patent field for the risk of missed judgment and the cost of misjudgment, the intervals were subdivided at a step size of 0.05, and ultimately 12 threshold combinations were formed. This not only ensured the granularity of parameter testing but also accurately located the optimal threshold range. Among multiple threshold combinations, combinations 1, 2, 5, and 9 performed particularly well in reducing duplicate similarity values and avoiding a high proportion of non similarity results, which met the optimization objectives. Among them, combination 1 showed the best performance in reducing duplicate values at 4%, while combination 9 demonstrated the highest measurement accuracy in actual patent content comparison, despite a high proportion of similarity of 0 at 46%. This discovery was consistent with Karim's team's research on efficient storage of SSN data, emphasizing the importance of threshold adjustment in improving data processing quality [22].

In summary, this study was the first to integrate the dynamic semantic perception of SSN with the deep text parsing of SAO. This effectively addressed the dynamic evolution of technical terms in the patent field and the heterogeneity of cross-source data formats. During the 500-second operation cycle, the processing time of this technology fluctuated slightly in the cluster environment, though this did not affect the overall performance. Compared to a single-machine environment, this technology was more stable and robust when handling data streams from multiple sensor nodes. However, there are still certain limitations in the research technology. At the sensor data level, if there is noise or incomplete sensor data, it will interfere with the initial integration quality of multi-source heterogeneous data. Noisy data may lead to misjudgment of semantic association, and incomplete data will cause information loss in SAO structure analysis. This affects the accuracy of subsequent similarity measurement.

4.2 Conclusion

Due to the different sources, formats, and structures of MSHD, integration is difficult. Therefore, a SSN-SAO-MSHD technology was proposed in this study. As the threads increased, the initial processing time of the system significantly decreased, reflecting the efficiency improvement brought by parallel processing. Regardless of the parallel processing strategy used, the system's processing time increased linearly with the number of sensor nodes. However, it still maintained good scalability overall, indicating that the SSN-SAO-MSHD was efficient and stable when processing large-scale data streams. SSN-SAO-MSHD had demonstrated higher stability and robustness in cluster environments, providing strong support for practical applications. In terms of similarity measurement, the study optimized the similarity measurement results by adjusting the SAO semantic structure matching threshold Q. As the threshold Q increased, the similarity repetition value and the proportion of patents with similarity of 0 both increase. To balance these two factors, the study analyzed the results of 12 different threshold settings and found that threshold combination 9 performed the best in reducing duplicate similarity values and the proportion of patents with similarity of 0. Meanwhile, its measurement results were also the most accurate. The limitation of the research is that it may not be comprehensive enough in optimizing performance in specific scenarios. Future research needs to further explore the impact of more variables on system performance to achieve widespread and efficient data integration and application.

Funding

The research is supported by: Research on Crime Prevention for Key Individuals Based on Big Data on the Internet; Project Number: SHZLYB2416.

References

[1] Di Chen, Fengbin Zhang, and Xinpeng Zhang. Heterogeneous IoT intrusion detection based on fusion word embedding deep transfer learning. IEEE Transactions on Industrial Informatics, 19(8):9183-9193, 2023.https://doi.org/10.1109/TII.2022.3227640

[2] Jade Ferreira, José Maria N. David, Regina Braga, Fernanda Campos, Victor Ströele, and Leonardo De Aguiar. Ontology-based data integration for the internet of things in a scientific software ecosystem. International Journal of Computer Applications in Technology, 67(2/3):256-262, 2022.https://doi.org/10.1504/IJCAT.2021.121533

[3] Luyi Bai, Nan Li, Lishuang Liu, and Xuesong Hao. Querying multi-source heterogeneous fuzzy spatiotemporal data. Journal of Intelligent and Fuzzy Systems, IOS Press, 40(5):9843-9854, 2021.https://doi.org/10.3233/JIFS-202357

[4] Oussama El Hajjamy, Hajar Khallouki, Larbi Alaoui, and Mohamed Bahaj. Semantic integration of traditional and heterogeneous data sources (UML, XML and RDB) in OWL2 triplestore. International Journal of Data Analysis Techniques and Strategies, 13(1-2):36-58, 2021.https://doi.org/10.1504/ijdats.2021.114667

[5] Anzhong Huang, and Fei Wu. Two-stage adaptive integration of multi-source heterogeneous data based on an improved random subspace and prediction of default risk of microcredit, Neural Computing and Applications, 33(9):4065-4075, 2021.https://doi.org/10.1007/s00521-020-05489-z

[6] R. Thirumahal, G. Sudha Sadasivam, and P. Shruti. Semantic integration of heterogeneous data sources using ontology-based domain knowledge modeling for early detection of COVID-19. SN Computer Science, 3(6):1-13, 2022.https://doi.org/10.1007/s42979-022-01298-4

[7] Marcos Zarate, Germán Braun, Mirtha Lewis, and Pablo R. Fillottrani. Observational/hydrographic data of the South Atlantic Ocean published as LOD, Semantic Web, 13(12):1-12, 2021.https://doi.org/10.3233/SW-210426

[8] Iman Naja, Milan Markovic, Peter Edwards, and Caitlin Cottrill. A semantic framework to support AI system accountability and audit. The Semantic Web, 160-176, 2021.https://doi.org/10.1007/978-3-030-77385-4_10

[9] Yuanyi Chen, Yih-Yeong Lin, Pengfei Yu, Yanyun Tao, and Zengwei Zheng. A graph-based sensor recommendation model in semantic sensor network. International Journal of Distributed Sensor Networks, 168536-168547, 2022.https://doi.org/10.1177/15501477211049307

[10] Magnus Palmblad, Enahoro Asein, Nina P Bergman, Arina Ivanova, Lukas Ramasauskas, Hazzar Mohammed Reyes, Stefan Ruchti, Leonardo Soto-Jácome, and Jonas Bergquist. Semantic annotation of experimental methods in analytical chemistry. Analytical Chemistry, 94(44):15464-15471, 2022.https://doi.org/10.1021/acs.analchem.2c03565

[11] Hebbi Chandravva, and Mamatha hr. Comprehensive dataset building and recognition of isolated handwritten Kannada characters using machine learning models. Artificial Intelligence and Applications, 1(3):179-190, 2023.https://doi.org/10.47852/bonviewAIA3202624

[12] Danila Vaganov, Egor Shikov, Anton Lysenko, and Polina Andreeva. Ontological model identification based on data from heterogeneous sources. Procedia Computer Science, 229:305-314, 2023. https://doi.org/10.1016/j.procs.2023.12.032

[13] Lei Ma, Yanning Zhang, and Vicente García-Díaz. Design and implementation of a fast integration method for multi-source data in high-speed network. Journal of High-Speed Networks, 29(3):251-263, 2023.https://doi.org/10.3233/JHS-222047

[14] Sushovan Das, Suman Bhowmik, and Chandan Giri. Cross-layer MAC protocol for semantic wireless sensor network. Wireless Personal Communications, 120:3135-3151, 2021.https://doi.org/10.1007/s11277-021-08603-z

[15] Xi Chen, Zhaoyang Yin, and Miaomiao Zhu. Semantic interaction strategy of multiagent system in large-scale intelligent sensor network environment. J. Sensors, 2022:1-10, 2022.https://doi.org/10.1155/2022/5969130

[16] Youngjoon Yu, Hong Joo Lee, Byeong Cheon Kim, Jung Uk Kim, and Yong Man Ro. Towards robust training of multi-sensor data fusion network against adversarial examples in semantic segmentation. IEEE International Conference on Acoustics, 4710-4714, 2021.https://doi.org/10.1109/ICASSP39728.2021.9413772

[17] Xue-Zhen Zhang, Bi-Hui Yu, and Chang Liu. SSN_SEM: Design and application of a fusion ontology in the field of medical equipment, Procedia Computer Science, 183:677-682, 2021.https://doi.org/10.1016/j.procs.2021.02.114

[18] Lidong Yang. Feature extraction of english semantic translation relying on graph regular knowledge recognition algorithm. Informatica, 47(8):103-124, 2023.https://doi.org/10.31449/inf.v47i8.4901

[19] Hadda Ben Elhadj, Farag Sallabi, Amira Henaien, Lamia Chaari, Khaled Shuaib, and Maryam Al Thawadi. Do-Care: A dynamic ontology reasoning

based healthcare monitoring system, Future Generation Computer Systems, 118:417-431, 2021.https://doi.org/10.1016/j.future.2021.01.001

[20] Bi-Hui Yu, He Wang, Xin-Peng Dong, and Xue-Zhen Zhang. Design and implementation of a semantic gateway based on SSN ontology. Procedia Computer Science, 183:432-439, 2021. https://doi.org/10.1016/j.procs.2021.02.081

[21] Mohamed Bettaz, and Mourad Maouche. Towards an ontological-based CIM modeling framework for IoT applications. Informatica, 48(4):663-684, 2024.https://doi.org/10.31449/inf.v48i4.5845

[22] Farah Karim, Maria-Esther Vidal, and Sören Auer. Compact representations for efficient storage of semantic sensor data. Journal of Intelligent Information Systems, 57:1-26, 2021.https://doi.org/10.1007/s10844-020-00628-3