Automation and Machine Learning (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/autml.2025.060115
ISSN 2516-5003 Vol. 6 Num. 1

Information Data Acquisition Method and Process Based on Artificial Intelligence

Haoyu Wang

Philippine Christian University Center for International Education, Manila, 1004, Philippines
haoyuwang1118@google.com

Keywords: Artificial Intelligence; Information Data; Acquisition Method; Process

Abstract: Data collection involves extracting, transforming, and standardizing data from various sources to facilitate subsequent analysis, processing, and utilization. This process includes multiple steps, such as identifying data sources and cleaning data. In the field of artificial intelligence, data collection is fundamental to achieving machine learning and deep learning technologies, enhancing prediction accuracy, and aiding in model training. This paper focuses on the methods and processes of artificial intelligence data collection, briefly explaining the principles, importance, and guidelines of AI data collection. It provides an in-depth analysis of different collection methods and emphasizes the specific procedures involved in data collection.

1. Introduction

Artificial intelligence (AI) is the fundamental theory and technology that explores how to use computer software and hardware to simulate certain intelligent behaviors of humans. Information data collection is the foundation for various activities, but manually collecting large volumes of information data is challenging and requires significant investment in human resources, materials, and time. To enhance the efficiency of information data collection, the application of AI should be strengthened to facilitate rapid data collection. Additionally, to ensure that the subsequent processing and use of information data are more convenient, the collection process should be strictly followed to improve the accuracy of the collected information and avoid issues such as duplicate data collection.

2. Content of artificial intelligence information data collection

2.1 Principle of artificial intelligence information data collection

The primary goal of AI data collection is to deeply mine and extract valuable information from the subject being tested, converting it into processable digital signals [1]. The principle involves detecting the subject to obtain its status information, such as using a temperature sensor to monitor environmental temperature. Based on this, the detected analog signals are converted into digital signals, a process that requires the assistance of a model converter to transform continuous analog signals into discrete digital signals. To improve the accuracy of the collected digital signals [2], methods such as filtering and amplification are applied. The processed signals are then transmitted to the data receiving end via media like optical fibers or wireless communication, and finally stored in memory, providing a reliable basis for subsequent analysis, processing, and decision-making, as illustrated in Figure 1.

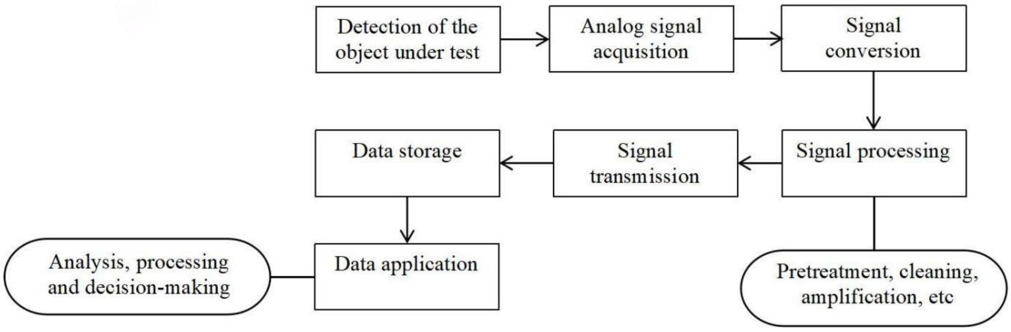
Detection of the object under test
Analog signal acquisition
Signal conversion
Signal processing
Signal transmission
Data storage
Data application
Analysis, processing and decision-making
Pretreatment, cleaning, amplification, etc

Figure 1: Principle diagram of artificial intelligence information data processing

2.2 Importance of artificial intelligence information data collection

Information data is the foundation of artificial intelligence technology. If the quality of information data is low, it will be challenging for AI systems to achieve accurate analysis and prediction [3]. The collection of information data involves multiple stages, including the selection of data sources and data storage. When selecting data sources, various methods are used for data collection, such as sensors and the Internet. Due to the differences in data sources, their formats and types often vary, so the selection should be based on the actual application scenario and requirements. Data acquisition, a crucial part of the data collection process, involves tasks like data scraping and cleaning, and must ensure that the entire process is legal and compliant, with a focus on protecting privacy. Data storage, which includes data organization and backup, should ensure that the data remains complete and usable [4].

2.3 Principles of artificial intelligence information data collection

In the data collection of artificial intelligence information, if you want to ensure that the information is true, accurate and compliant, you need to strictly follow the following principles, as shown in Figure 2.

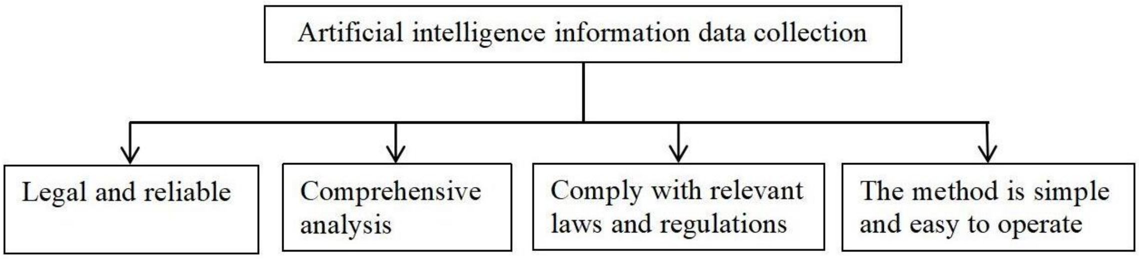
Artificial intelligence information data collection
Legal and reliable
Comprehensive analysis
Comply with relevant laws and regulations
The method is simple and easy to operate

Figure 2: Principles of artificial intelligence information data collection

(1) The data source should be legal and reliable, and all information data should be guaranteed to be true and have a high degree of credibility, so as to provide reliable data support for subsequent analysis and decision-making.

(2) During the period of information data collection, the diversity and representativeness of information data are comprehensively considered to improve the differentiation ability of the model [5].

(3) For information data collection, based on relevant laws and regulations, especially when involving personal privacy and sensitive information, strictly follow the data protection regulations to avoid the infringement of users' personal privacy and ensure the security of information.

(4) In the process of collecting information data, the methods are simple and convenient to operate, so that the speed of collecting information data is accelerated, and the manpower and time cost are controlled in the minimum range [6].

3. Artificial intelligence information data acquisition method

At present, there are many methods of artificial intelligence information data collection, including crawler capture, sensor, etc. Each method has its own advantages and disadvantages, and the specific way to adopt depends on the actual situation [7].

3.1 Crawler capture

During the process of information data collection, web crawlers are used to gather data from the Internet. Currently, Python web crawlers are widely used to extract data from websites, including news articles, comments, and user data. A web crawler is an automated program that primarily collects web content through HTTP requests and web page interactions, and parses the data information [8]. During data collection, this method involves four main steps: sending requests, receiving responses, parsing data, and storing data. Throughout this process, the crawler first sends HTTP requests to the target website, then promptly parses the returned data, quickly extracts the required information, and finally stores the data in a database or other locations. The specific process is illustrated in Figure 3.

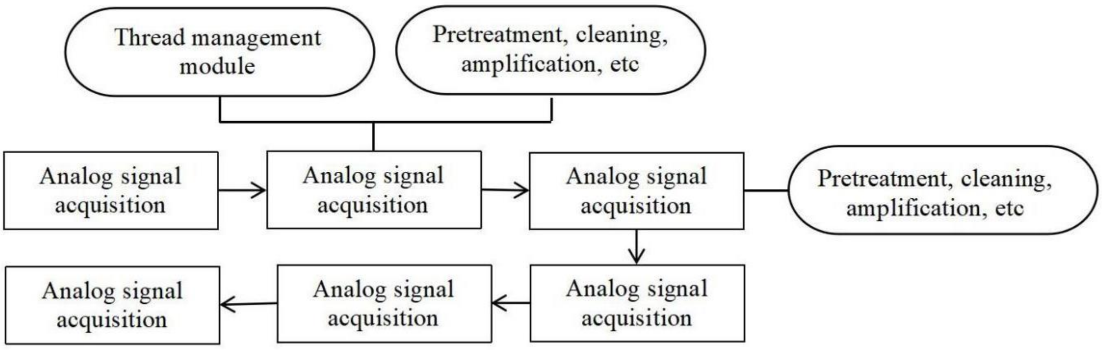
Thread management module
Pretreatment, cleaning, amplification, etc
Analog signal acquisition
Analog signal acquisition
Analog signal acquisition
Pretreatment, cleaning, amplification, etc
Analog signal acquisition
Analog signal acquisition
Analog signal acquisition

Figure 3: Data acquisition and processing process based on Python web crawler

3.2 Sensor acquisition

In the process of collecting artificial intelligence data, sensors can also be used to quickly gather data. Currently, there are various types of sensor devices, such as cameras, temperature sensors, and pressure sensors, which can collect real-time data based on actual conditions. The information collected by these sensors is then uploaded to the cloud via wireless transmission technology, enabling relevant personnel to conduct in-depth analysis and prediction [9].

3.3 Manual annotation

In the process of AI data collection, manual annotation can be used to process the data. This method involves manually converting raw data into structured data that machine learning models can understand and learn from. When annotating, the data is categorized and labeled based on specific task requirements. For example, in image recognition, the target object can be highlighted in the image, and its category accurately labeled. The quality of manual annotations is closely linked to the performance of AI models; high-quality annotations can improve model accuracy, enabling the model to accurately learn patterns and characteristics from the data and enhance its generalization ability. However, this method has certain drawbacks, requiring a significant investment of time and manpower, especially for large-scale data. The cost of annotation is high, and it is challenging to achieve the expected consistency and accuracy of annotations. There may be differences in how annotators interpret the same data, leading to variations in annotation results [10]. Additionally, the professional level and experience of the annotators can affect the quality of the annotations. To address these issues, auxiliary tools and methods, such as machine learning models, should be utilized to accurately annotate data, and manual review and correction of the annotations should be conducted to reduce the workload.

4. Information data acquisition process based on artificial intelligence

4.1 Requirement analysis and goal setting

In the early stage of artificial intelligence information data collection, it is necessary to conduct in-depth analysis on the objectives and requirements of the project, understand the types, amount and frequency of data collection, etc. At the same time, in-depth communication and exchange with all relevant parties should be carried out to ensure that the collection objectives are consistent with the project requirements [11].

4.2 Data source selection

Based on the results of the demand analysis, select data sources that are suitable for the requirements. This includes public datasets, commercial data, social media data, and sensor data. When selecting data sources, it is essential to reasonably assess their quality, clearly define their availability and acquisition costs, and ensure that the chosen data sources align with the requirements [12].

4.3 Design of data acquisition method

In this study, the main method of intelligent sensor data acquisition is based on the design of a data acquisition system to collect data information in an all-round way, covering data processing centers and sensor networks. The specific framework is shown in Figure 4.

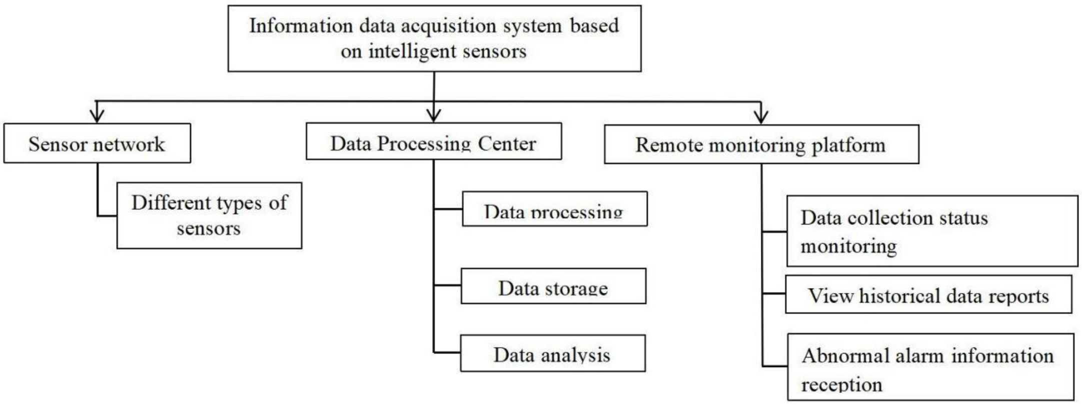
Information data acquisition system based on intelligent sensors
Sensor network
Different types of sensors
Data Processing Center
Data processing
Data storage
Data analysis
Remote monitoring platform
Data collection status monitoring
View historical data reports
Abnormal alarm information reception

Figure 4: Framework of information data acquisition system based on intelligent sensor

(1) The components of the sensor network are all kinds of high-precision intelligent sensors, which collect important data information in an all-round way. In the process of connection, sensors use wired or wireless methods to ensure the accuracy of the collected information data.

(2) The main function of the data processing center is to comprehensively receive and process the data of the sensor network. With the help of data processing algorithms, valuable information data can be extracted to provide reliable reference for subsequent decision-making.

(3) The remote monitoring platform provides a convenient human-computer interface, remote access to the data processing center, dynamic grasp of data collection, historical data reports, real-time monitoring and receiving abnormal alarm information.

4.4 Sensor selection and configuration

Considering the specific framework and requirements of the system, the selection and configuration of sensors are crucial. Based on specific application scenarios and needs, it is essential to choose intelligent sensors with high performance, stability, and accuracy to ensure more accurate and reliable data collection [13]. When evaluating the parameters of the sensors, such as sensitivity, response time, and operating range, they should align with the system's specific requirements. Restructured to use the active verb "design". Uses "and prevent" for parallelism. "Rationally" modifies "design" more directly. The sensor configuration should be compatible with the data acquisition system and control center, ensuring high integration and unified management.

4.5 Information data collection

4.5.1 Analog signal acquisition

During the data acquisition phase, it is essential to efficiently collect and process analog signals. In this process, the original time interval is used as a benchmark to handle the continuously changing analog signals and obtain sample values. To ensure a smoother conversion of these sample values, hold technology is employed to maintain the signal's stability during the conversion process. The A/D converter is used to convert the signal samples into digital signals, facilitating subsequent data processing and analysis. It is important that the converter used in this process maintains high performance and accuracy to ensure precise [14] conversion results.

4.5.2 Digital signal acquisition

Digital signal acquisition can accurately capture and record information data, characterized by high resolution. During the data acquisition process, this technology minimizes errors, thereby enhancing the system's overall performance. Additionally, digital signals have strong resistance to interference, effectively resisting electromagnetic interference during data transmission and processing, ensuring data stability.

4.5.3 Wireless transmission

During the data collection process, wireless transmission technology is utilized to enhance efficiency and make the entire process more flexible. Currently, Bluetooth, Wi-Fi, and LoRa are the most widely used wireless transmission technologies. Each technology has its own advantages and disadvantages, and the appropriate technology should be chosen based on specific needs. Wi-Fi offers stable and efficient data transmission. If the data volume is large and timeliness is critical, Wi-Fi can be used for rapid data transmission to the cloud or data center. Bluetooth technology is known for its short transmission range and low power consumption, making it ideal for short-range communication and enabling instant data transmission and sharing. LoRa technology, with its strong communication capabilities and low power consumption, can transmit data over long distances. LoRa sensors, which operate on the LoRa network, upload data to remote servers at high speed, facilitating rapid data collection and monitoring.

4.6 Information data preprocessing

In the process of collecting AI data, data processing is a crucial step that directly impacts the accuracy of the collected information, and thus should be given due attention. In the data processing phase, data filtering and nonlinear correction are particularly important. These steps not only enhance data quality but also provide a reliable foundation for subsequent analysis, ensuring the accuracy of the results.

Data filtering, a core part of preprocessing, primarily employs algorithms such as moving average filtering to remove noise and outliers from the data. This process enhances the smoothness of the information data curve, leading to more accurate data acquisition. This step significantly impacts subsequent data analysis and processing, preventing errors and misjudgments caused by excessive data fluctuations, thereby enhancing the accuracy and stability of the analysis results.

(2) Nonlinear correction. Based on the sensor's output characteristics and the ideal linear relationship, the existing deviations are precisely and flexibly adjusted. Since the data obtained by sensors often fails to maintain perfect linearity during output, software algorithms are used to assist in quickly completing nonlinear correction. This helps to compensate for hardware design flaws, thereby optimizing system performance. This process ensures that measurement results closely match the actual values, providing a reliable data reference for subsequent decision-making and control measures.

4.7 Data cleaning

The primary goal of data cleaning is to address and correct issues such as noise, errors, duplicates, and missing values in the collected raw data, thereby enhancing the quality and usability of the data. During the data cleaning process, a thorough initial inspection of the data is conducted to comprehensively scan and evaluate the collected data, identify any issues, and address them. For instance, when dealing with missing values, methods such as deletion or filling are employed based on the data's nature and the extent of the missingness, as detailed in Table 1.

Table 1: Data cleaning problems and processing

order number	question	handle
1	Missing values	Delete records, fill in (mean, median, etc.)
2	duplicate record	Delete and merge duplicate records
3	erroneous data	Delete and amend
4	outlier	Delete, correct, and mark
5	noisy data	Filter, remove outliers, remove irrelevant characters

4.8 Data storage

In terms of information data collection, data storage is the key. At present, there are two forms of data storage: local storage and cloud storage, each with its own unique advantages and disadvantages:

(1) Local storage. This method involves storing data directly within the sensor or the device that collects it, offering high data access efficiency and minimal reliance on network conditions. It is

suitable for scenarios with high real-time requirements or in unstable network environments. However, the limited capacity of local storage makes it challenging to store large volumes of data for extended periods, and the data lacks robust security measures, posing risks such as data loss and corruption.

(2) Cloud storage model. This method involves uploading data to remote servers, enabling centralized management of information and data, enhancing scalability, and improving data security and reliability. Cloud storage offers robust data backup capabilities and high recovery levels, facilitating data sharing and collaboration. With the widespread adoption of cloud computing technology, the cost of cloud storage is continuously decreasing, making it highly effective for large-scale data storage. However, this approach heavily relies on network conditions, and real-time performance can be affected by network latency and other factors. Additionally, reasonable security measures must be implemented to protect data privacy.

Regardless of the method used, during the specific storage process, it is essential to categorize data based on its type and source, to facilitate future queries and usage. Additionally, regular backups should be made to ensure data security and enhance its recoverability. Furthermore, based on project requirements, manage data access permissions to ensure that only authorized personnel can access sensitive data.

4.9 Feedback and improvement

During the data collection process, it is essential to establish and improve feedback and improvement mechanisms. Especially when using the data, it is crucial to gather comprehensive user feedback to identify the data's effectiveness and any existing issues. Regular evaluation of the data collection process shall be performed to identify problems and implement targeted improvements. Feedback and evaluation outcomes shall subsequently be used to iteratively optimize the process for greater adaptability.

4.10 Result analysis

Taking the data collection from warehouse shelves as an example, this project primarily employs intelligent sensors and RFID technology, along with big data analysis and machine learning algorithms, to automatically collect, process, and analyze information data, thereby enhancing enterprise efficiency. The specific results are shown in Table 2. According to the data in the table, when collecting and processing information data, the methods mentioned above are strictly followed, achieving a high accuracy rate of 99.8% in data reading, surpassing industry standards. The reliability of sensor data reaches 99.7%, also exceeding industry standards. Overall, by leveraging advanced technical means for data collection, the process is not only well-structured but also highly efficient and accurate.

Table 2: Comparison of intelligent data acquisition test results

order number	performance index	test result	occupation standard
1	Data read accuracy	99.8%	98.5%
2	Sensor data reliability	99.7%	99.0%
3	Average response time	40ms	200ms
4	Accuracy of prediction	96.8%	85%
5	Energy efficiency improved	35%	20%
6	Labor cost savings	45%	25%

5. Conclusion

Overall, the methods and processes for collecting artificial intelligence information data are characterized by their complexity and systematic nature, involving multiple stages such as requirement analysis, data processing, and data storage. To ensure the accuracy of the collected data, it is essential to establish a comprehensive data collection process that guarantees data quality, ensuring reliable data references for subsequent model training and application. As technology advances, data collection tools and methods will continue to innovate, and professionals should focus on staying updated with new technologies to enhance their competitiveness in an ever-changing environment.

References

[1] Zhao Yinhao, Wang Yalin, Qin Guangtao, et al. Research and Application of Automatic Data Acquisition and Intelligent Processing Technology for Experiments [J]. Computer Knowledge and Technology, 2025,21(15):82-86.

[2] Chen Haigen, Dong Xinying, Jin Jiangtao, et al. Design and Implementation of a Full-Station Survey Data Acquisition Program for Android/Hongmeng Smart Terminals [J]. Urban Survey, 2025, (02):204-208.

[3] Wang Lei. Research and Application of Intelligent Collection System for Financial News Data Set [J]. Industry and Information Technology Finance Science and Technology, 2025, (02):48-59.

[4] He Jing. Research on mechatronics data acquisition method based on intelligent sensors [J]. Science and Technology Innovation, 2025, (07):85-88.

[5] Gao Chen, Ye Baozhu, Liu Haidong, et al. Application of Intelligent Computing in Data Collection and Analysis for Electricity Consumption Inspection [J]. Integrated Circuit Applications, 2025,42(04):302-303.

[6] Chen Jinquan. Research on Data Acquisition and Management Methods for Incremental Distribution Networks Based on Intelligent Communication Protocols [J]. Communication Power Technology, 2025,42(3):43-45.

[7] Yan Chong. AI-based automated data collection technology for digital media [J]. Radio and Television Network, 2025,32(03):39-42.

[8] Yang Chengzhi. Research on the Real-time Data Acquisition and Analysis System for Intelligent Storage Racks [J]. China Machinery, 2025, (08):102-105.

[9] JahaniRahaei A ,Milelli M ,Chiesa G . Urban weather dataset for building energy simulations: Data collection and EPW file generation for Torino, Italy (2014–2023) [J]. Data in Brief, 2025, 61 111708.

[10] Cian H ,Dou R ,Irwin C . Embedding historical and contextual sensitivity in QuantCrit approaches to STEM identity research: implications for data collection and analysis techniques [J]. Current Opinion in Behavioral Sciences, 2025, 64 101530.

[11] Antonova P D ,Jelyazkov J ,Pavlova I . Air quality monitoring platform with multiple data source support [J]. Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, 2025, 47 (1): 8454-8470.

[12] Klein K ,Muller A ,Wohde A , et al. An AI-assisted workflow for object detection and data collection from archaeological catalogues [J]. Journal of Archaeological Science, 2025, 179 106244.

[13] Rahman A A ,Mridul C M ,Roy P , et al. A Multi-Head Attention mechanism assisted MADDPG algorithm for real-time data collection in Internet of Drones [J]. Vehicular Communications, 2025, 54 100944.

[14] Nikooharf H M ,Shirinbayan M ,Ghodsian N , et al. Toward advance/digitalized FFF: real-time multimodal synchronized data acquisition and ML/DL-driven process optimization [J]. Progress in Additive Manufacturing, 2025, (prepublish): 1-18.