

Article

Identification of Environmental Pollutants in Construction Site Monitoring Using Association Rule Mining and Ontology-Based Reasoning

Zhao Xu , Huixiu Huo  and Shuhui Pang 

Department of Civil Engineering, Southeast University, Nanjing 211189, China

* Correspondence: xuzhao@seu.edu.cn

Abstract: Pollutants from construction activities of building projects can have serious negative impacts on the natural environment and human health. Carrying out monitoring of environmental pollutants during the construction period can effectively mitigate environmental problems caused by construction activities and achieve sustainable development of the construction industry. However, the current environmental monitoring method relying only on various sensors is relatively singular which is unable to cope with a complex on-site environment. We propose a mechanism for environmental pollutants identification combining association rule mining and ontology-based reasoning and using random forest algorithm to improve the accuracy of identification. Firstly, the ontology model of environmental pollutants monitoring indicator in the construction site is built in order to integrate and share the relative knowledge. Secondly, the improved Apriori algorithm with added subjective and objective constraints is used for association rule mining among environmental pollutants monitoring indicators, and the random forest algorithm is applied to further filter the strong association rules. Finally, the ontology database and rule database are loaded into a Jena reasoning machine for inference to establish an identification mechanism of environmental pollutants. The results of running on a real estate development project in Jiangning District, Nanjing, prove that this identification mechanism can effectively tap the potential knowledge in the field of environmental pollutants monitoring, explore the relationship between environmental pollutants monitoring indicators and then overcome the shortcomings of traditional monitoring methods that only rely on sensors to provide new ideas and methods for making intelligent decisions on environmental pollutants in a construction site.

Keywords: environmental pollutants; monitoring system; association rule mining; ontology-based reasoning; random forest; construction site



Citation: Xu, Z.; Huo, H.; Pang, S. Identification of Environmental Pollutants in Construction Site Monitoring Using Association Rule Mining and Ontology-Based Reasoning. *Buildings* **2022**, *12*, 2111. <https://doi.org/10.3390/buildings12122111>

Academic Editors: Xiao Li, Yue Teng, Chengke Wu, Hengqin Wu and Geoffrey Qiping Shen

Received: 3 November 2022

Accepted: 28 November 2022

Published: 1 December 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The rapid development of the construction industry has made significant contributions to the improvement of China's national economy and people's lives, but it has also caused a relatively obvious negative impact on the natural environment [1–3]. Compared with other projects in the construction industry (such as bridge projects or tunnel projects), construction projects are more concentrated in places where human beings gather, which will have a more serious influence on the daily life and physical health of the surrounding people [4–7]. According to statistics, environmental pollution caused by various types of construction activities accounts for 34% of all pollution, and this is especially dominated by the environmental problems generated during the construction phase of the project [8]. Therefore, the management of environmental pollutants from construction sites is essential to reduce the impact of construction on human beings and to further realize the sustainable development of the construction industry.

The focus of previous studies on the control of environmental pollutants at construction sites can be divided into three parts: pollutants prediction in the pre-construction

phase, quantitative analysis in the post-construction phase and real-time monitoring in the construction phase. However, both the pollutants prediction before construction and the quantitative analysis after construction have some unavoidable defects, which are specifically reflected in the following two points: prior studies [9,10] on pollutants prediction fail to predict pollutant emissions where unpredictable and complex labor is undertaken; quantitative analysis [11–15] using experimental methods in the post construction stage cannot effectively identify the potential emission risks in the construction process. These functions depend on the real-time monitoring of environmental pollutants in the construction phase to achieve accuracy. At present, in the context of green construction and smart site development in the construction industry, the construction site of each project has basically realized the full configuration of environmental monitoring equipment, mainly using various types of sensors to sense the state of the environment in real time, and when the index monitoring value exceeds the system preset value, the recognition signal is issued. Further, in order to achieve more accurate and timely monitoring at construction sites, Internet of Things technology [16–18], sensor networks [19], convolutional neural network [20,21] and some software programs [22] are introduced. However, the traditional pollutants identification mechanism is relatively simple [23], when the monitoring data collected and transmitted are abnormal or real environmental emergencies occur, the management personnel are likely to fail to make the correct judgment in time due to the lack of relevant professional knowledge. Besides, there are still another limitation of the previous research: the field of environmental pollutants monitoring involves a wide range of knowledge, and the research topics are cultivated under the joint action of engineering technology and environmental ecology [24], requiring multi-disciplinary synergy, while the current knowledge distribution is fragmented and usually presented in the form of written text or expert experience, making it difficult to realize the integration and sharing of knowledge. There is thus a need to establish a way to identify environmental pollutants in construction site monitoring so that on the one hand knowledge in the field can be integrated and shared, and on the other hand when abnormal monitoring data and unexpected accidents occur with traditional monitoring methods, further judgments on the current environmental pollution situation can be made based on the identified pollutants, improving monitoring efficiency and accuracy.

Association rule mining (ARM) is a knowledge discovery method for finding correlations and causal associations between frequent items or attribute sets in a database, first proposed by Agrawal et al. [25] in 1993 for the shopping basket problem, and is mainly used to discover hidden association relationships between data. In the domain of construction engineering, the association rule is frequently employed to uncover correlations between causes and accidents [26] and risk monitoring [27] in the construction safety area to compensate for the weaknesses of qualitative approaches in finding the probable causes of construction accidents. In essence, ARM is a simple and useful approach for identifying frequent itemsets between unknown parameters and generating strong association rules from huge datasets, especially since it can find frequent itemsets and association rules between diverse anomalous circumstances that have been tracked [28]. Therefore, association rule mining can be adopted to analyze the association relationship between various indicators of environmental pollutants. When the sensor at the construction site fails or the data transmission is abnormal, the association rule can be implemented to jointly identify the risk factors using other related parameters, thus reducing, to a certain extent, the errors caused by the insufficient amount of sensor information. Using association rule mining alone cannot build a readable and verifiable model [29], and the strong association rules obtained from association rule mining can be further verified utilizing the random forest algorithm to improve the accuracy rate. Conventional data analysis methods are not easily avoiding the interaction between independent variables, while random forest, as one of the emerging machine learning algorithms, requires low datasets, runs stably and has no overfitting and covariance problems [30]. Many scholars have studied the combination of association rules and random forests, all of which have shown excellent accuracy and

advantages in addition. Lee et al. [31] used the random forest algorithm for the post-mining process of association rules and based on this process, mines RME terms from the text to further construct domain knowledge. Qu et al. [32] proposed a joint association rule and random forest regression algorithm for seismic multi-attribute sand body thickness prediction and numerically verified that the method can effectively identify redundant features in seismic multiple attributes.

Of necessity, the knowledge in the field of environmental pollutants monitoring at construction sites requires integration and sharing before using association rule mining analysis, and this function needs to be realized by building ontology models. Ontology is the formal conceptualization of knowledge in a certain domain [33] which helps ensure consistent semantics and avoid inconsistent representations under different systems. Ontology-based reasoning is based on the concepts and properties of ontologies and extracts the knowledge implicit in explicit definitions and statements through a processing mechanism. In the field of construction engineering, the technology of ontology-based reasoning has also been deployed and promoted, including domain knowledge base establishment, construction safety control and project risk prevention and control. However, there are certain shortcomings in this area of research. First, few ontology models for environmental pollutants monitoring have been proposed; second, the rule base used to accomplish ontology-based rule inference is either predefined [34] or based on expert opinions [35], which lacks constant updates. In contrast, using strong association rules derived by association rule mining methods as the rule base for ontology reasoning can nicely solve the above problems. The rule base thus formed is not static but is continuously learned through association rules. In summary, the purpose of this paper is to construct an environmental pollutant identification mechanism, founded on association rule mining and ontology-based reasoning, to explore the relationship between environmental pollutant monitoring indicators and then to overcome the shortcomings of traditional monitoring methods relying only on sensors.

The remainder of the paper is structured as follows: Section 2 presents related work. Section 3 describes the methods used in the study including ontology, association rule mining, Jena reasoning rules, random forest and identification mechanism. Section 4 demonstrates the concrete implementation of the method using a specific real estate development project in Jiangning District, Nanjing. Finally, the paper is concluded along with future work.

2. Literature Review

2.1. Monitoring of Environmental Pollutants

In the context of mutual integration of industrialization and informatization, information technology, such as Internet of Things (IOT), Building Information Model (BIM), Remote Sensing (RS), Global Position System (GPS) and Geographic Information System (GIS), plays a role in monitoring environmental pollutants at construction sites, helping to realize real-time perception of environmental conditions at construction sites and developing multifunctional monitoring system platforms. Wong et al. [36] developed an integrated microenvironmental system which, combined with the application of Global Position System (GPS), Wireless Fidelity (WIFI), 3rd-Generation (3G) and other technologies, can transmit the monitoring information of a location to the system in real time for analysis and evaluation by the managers. Smaoui et al. [37] created a real-time monitoring system that used a low-cost dust sensor to track worker dust exposure and displayed the data using BIM. Kim et al. [22] created a system (CPSM) to control particulate matters in real time at different locations on a building site using Internet of Things technology.

Most of the previous researchers started from improving the monitoring technology of environmental pollutants at construction sites to improve its sensitivity and precision. Even so, there are still many defects in the monitoring technology, and the accuracy of environmental pollutants monitoring cannot be firmly guaranteed when there are abnormalities in the monitoring data collected and transmitted at construction sites. In addition,

the utilization rate of construction site environmental monitoring data is low, and there is a lack of a mechanism for knowledge sharing.

2.2. Association Rule Mining

Knowledge Discovery in Database (KDD) research becomes very active due to the wide application of relational databases and the advantages of a unified organizational structure, an integrated query language and equality between relations and attributes [38]. The term was first proposed by Piatetsky-Shapiro at the International Joint Academic Conference on Artificial Intelligence held in 1989 and was classically defined by scholars such as Fayyad and Piatetsky-Shapiro in 1996: knowledge discovery refers to the non-trivial identification of valid, novel, potentially useful and ultimately understandable patterns from raw data process.

Association rule mining is an indispensable tool to realize the process of knowledge discovery that is used to find interesting relationships between variables in huge databases. At present, the research related to association rule mining is more mature, and it is generally solved using exact algorithms or intelligent algorithms, such as Apriori algorithm [25], FP-growth (frequent pattern growth) algorithm [39], Eclat algorithm [40], genetic algorithm [41,42], ant colony optimization algorithm [43] and particle swarm Optimization algorithms [44]. The Apriori algorithm is the most basic and widely used algorithm in discovering the association rules among different attributes, which has the advantage of simple ideas and easy implementation.

In the field of construction engineering, association rule mining has also been widely studied in applications. Cabello et al. [26] utilized the association rule method of data mining to extract knowledge from the historical data of construction accidents to identify potential hazards to develop effective safety procedures. Sun et al. [45] classified the window-opening duration modes in the observed workplace using the K-means clustering approach. Then, association rule mining was used to discover the window-opening behavior modes and contributing variables. Tao et al. [46] applied the method of association rules to explore the causes of collapse in construction accidents. Antonio et al. [26] utilized the association rule approach to extract information from past occupational data of construction accidents in order to assist managers in identifying common situations that may be avoided in the future. Association rule mining algorithms can be very good at analyzing the correlation of data, especially the cause-effect relationship between data, and, therefore, can fully exploit the invisible knowledge in the domain. At present, the application of association rule mining in the construction engineering is mostly focused on energy consumption and safety management, and it has not been studied in depth in the monitoring of environmental pollutants.

2.3. Ontology-Based Reasoning

The concept of ontology has been gradually enriched since it was introduced into the computer field in the 1980s and 1990s and applied to artificial intelligence. Gruber [33] provided a classical definition of ontology as a normative account of conceptualization, but this definition does not fully summarize the essence of ontology. Borst et al. [47] added more to clarify the nature of ontology and proposed the concept of ontology sharing. Studer et al. [48] continued to expand on Borst's definition by proposing a definition that is highly accepted by expert scholars in various fields: an ontology is an explicit formal specification of a shared conceptual model.

With the deepening of ontology research, scholars have constructed a large number of ontologies and believe that these ontologies can be classified into four major categories based on different characteristics. Table 1 compares the different ontology types.

In recent years, with the deepening research on ontology theory, ontology-based reasoning has been broadly applied in the fields of knowledge management [35], semantic segmentation [53], biomedicine [54] and risk warning [55]. In the field of construction engineering, this technology has also been deployed and promoted, including domain

knowledge in base construction, construction safety control and project risk prevention. Tserng et al. [56] developed an ontology-based risk management framework that can improve the effectiveness of risk management using past experience. Moradi et al. [57] formulated a conceptual model of quality assurance ontology that can provide a unified representation of information for consistency testing of urban construction quality. Based on the above analysis, reasoning about ontologies has multiple applications: for the builders of ontologies, the main role of reasoning is to detect conflicts and optimize representations and ontology fusion; for the users of ontologies, the role of reasoning is mainly to obtain knowledge in ontologies and apply knowledge in ontologies to solve problems.

Table 1. Classification of ontologies.

Ontology Type	Characteristics
Representation ontology [49]	It is related to knowledge representation and refers to the ontology used to obtain the expression of meta-words that formalize knowledge in a particular knowledge representation system.
Generic ontology [50]	It is mainly used to study generic concepts and relationships between concepts, independent of a specific domain, and can be shared on a larger scale.
Domain ontology [51]	It focuses on concepts and the relationships between concepts in a specific subject area and is a specialized ontology.
Application ontology [51,52]	It describes knowledge that relies on both a specific field and a topic and is linked to domain-specific expertise and problem-solving methods.

3. Methodology

Aiming at the problems of difficult integration and sharing of knowledge in the process of environmental pollutants monitoring in construction site, the ontology model is firstly constructed on the premise of ontology theory. Secondly, the classical algorithm of association rules, Apriori algorithm, is improved by adding subjective and objective interest constraints to form the constrained Apriori algorithm to address the shortcomings of low mining efficiency of traditional algorithm. The random forest algorithm is applied to further filter the strong association rules. Finally, on the basis of the ontology model and association rule mining results, Jena rule language is used to write reasoning rules in order to establish an identification mechanism of environmental pollutants. The overall methodology is presented in Figure 1.

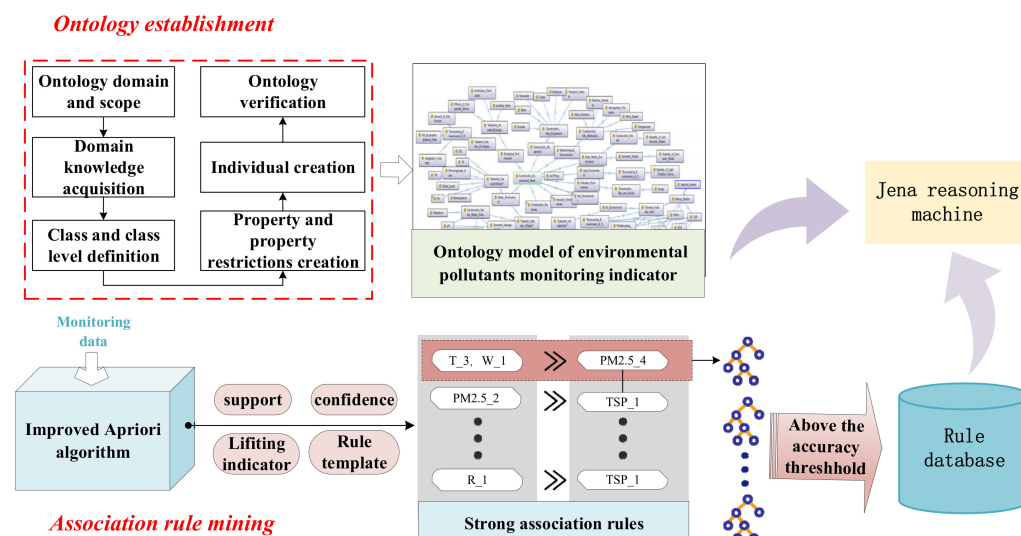


Figure 1. The methodology of this research.

3.1. Ontology Establishment

A complete ontology usually contains five basic elements: Concept, Relations, Function, Axioms and Instances, which are expressed as:

$$O = \{C, R, F, A, I\} \quad (1)$$

In the above formula:

O—Complete ontology model;

C—Concept, also called Class, is a model for describing objects or events in a domain;

R—Relations, used for describing the links between concepts or classes;

F—Function, a special representation of relations between concepts, connects two concepts with a mapping relation;

A—Axioms, a recognized fact or rule in the ontology, is used to constrain classes and relationships;

I—Instances, representing the physical presence of a certain class, as concrete objects of this class.

There is no unified standard for an ontology construction method at present. Based on different application fields and requirements, scholars have proposed skeleton method [58], TOVE method [58], IDEF5 method [59], METHONTOLOGY method [60], five-step cycle method [61] and seven-step method [59]. Among them, the application of Stanford's seven-step method is relatively mature, which is dedicated to the construction of domain ontology. The semi-automatic construction of ontology can be realized by using Protégé software, and the constructed ontology is highly detailed. This paper will use the seven-step method to establish the ontology.

(1) Ontology domain and scope

The ontology model of environmental pollutants monitoring indicator in construction site belongs to the field of construction engineering, and the research scope is environmental pollution and ecological destruction in the construction stage of the project.

(2) Domain knowledge acquisition

It is vital to obtain the existing ontology and important terms related to “construction engineering”, “project construction” and “environmental pollutants monitoring” and to acquire the knowledge needed to construct ontology by consulting the literature, standards and expert research.

The available ontology is searched from top ontology, database, online ontology library, saurus and other ontology resource libraries. The existing ontology related to this ontology model in the search results includes green construction ontology, construction quality ontology, construction noise ontology and safety construction ontology. However, these ontologies are different from the research direction of this paper, and only parts of the contents are similar and cannot be reused directly. Therefore, this paper will obtain the domain-important terminology to reconstruct a new ontology based on these ontologies.

Firstly, important terms in the field are extracted from relevant national standards, books, papers and internet information. Secondly, for the preliminary term set, it is required to listen to the opinions of relevant experts, standardize term expression, delete duplicate concepts and ensure the uniqueness of semantics. As a consequence, the important terms in the relevant field will be represented as classes or properties in the ontology.

(3) Class and class level definition

Class is the most fundamental component unit in ontology, which is an abstract description of objective things with the same characteristics. On the level classification of classes, it generally includes three methods: top-down method, bottom-up method and comprehensive method. Top-down method starts from top-level concepts, refines decomposition step by step and finally forms concrete entities; bottom-up method indicates that the

top-level concept is formed by starting from specific examples and gradually summarizing; comprehensive method is the comprehensive application of the above two methods.

In this paper, the comprehensive method is used to define the class level, and the concepts in the relevant field are preliminarily abstracted into nine categories: Construction Project, Environment Impact, Monitoring Item, Monitoring Indicator, Monitoring Site, Monitoring Time, Monitoring Level, State of Environment and Countermeasure, as shown in Figure 2.

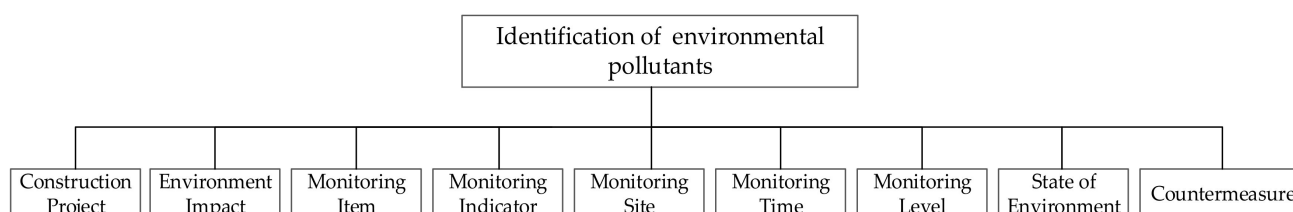


Figure 2. Basic classes of the ontology model for identifying environmental pollutants.

(4) Property and property restrictions creation

The next task is to describe the relationship between classes and the characteristics of the class itself, that is, the definition phase of object properties and data properties. At the same time, in order to construct the ontology model more accurately and completely, we are required to define the Domains, Ranges and Characteristics of properties and add Quantifier Restrictions, Cardinality Restrictions or “has” Value Restrictions. Some object properties and data properties are shown in Tables 2 and 3.

Table 2. Definition of object properties.

Object Properties	Domains	Ranges
has Environment Impact	Construction Project	Environment Impact
has Monitoring Item	Construction Project	Monitoring Item
has Monitoring Indicator	Monitoring Item	Monitoring Indicator
has Monitoring Site	Monitoring Indicator	Monitoring Site
is Monitored in	Monitoring Indicator	Monitoring Time

Table 3. Definition of data properties.

Data Properties	Domains	Ranges
has Project Name	Construction Project	xsd: string
has Project Scale	Construction Project	xsd: string
has Construction Time	Construction Project	xsd: string
has Monitoring Value	Monitoring Item	xsd: decimal
has Monitoring Unit	Monitoring Item	xsd: string

(5) Individual creation

Individual creation is a crucial step in the process of ontology model construction, which realizes the correspondence between abstract concepts and specific examples and enriches the ontology knowledge database. Specific individuals in this paper will be illustrated in Section 4.

(6) Ontology verification

After the completion of the individual creation, the whole ontology model has been preliminarily constructed. However, due to semantic contradictions or logical errors in the

ontology creation process, it is still important to check the consistency of the initial ontology model. In this paper, the HermiT 1.4.3.456 reasoning machine of Protégé software is used for reasoning to complete ontology verification, including checking the consistency of ontology and data description based on OWL language to judge the consistency of statement logic; according to the relationship between model declarative reasoning concepts, judging whether the declarative logic conforms to the domain understanding; and identifying repeated names. Errors are corrected in the ontology model until the consistency check is passed, and the final ontology model is shown in Figure 3.

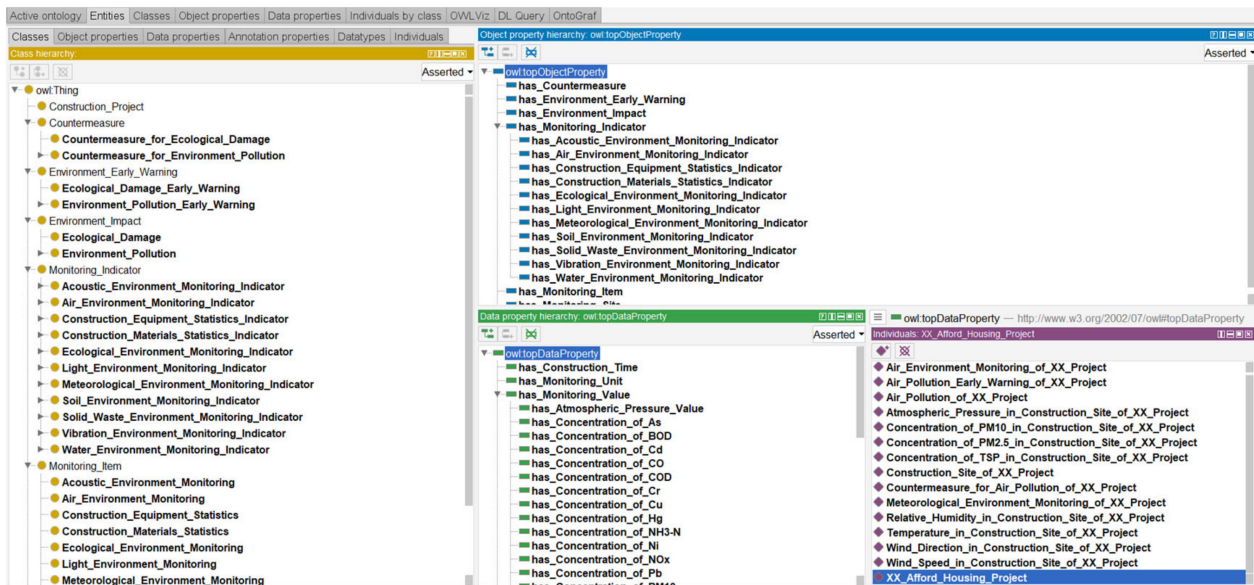


Figure 3. Final ontology model.

3.2. Association Rule Mining

Association rule is a rule-based machine learning method to reflect the interdependence or relevance between one thing and other things [62]. The typical application of association rules is to analyze the shopping basket data in supermarkets and analyze customers' purchasing habits by discovering the relationship between different commodities in the shopping basket. The basic model of association rules mainly includes the concepts of itemset, frequent itemset, support number, support degree and confidence degree, which are introduced as follows:

Suppose $I = \{I_1, I_2, \dots, I_m\}$ is a collection of all items, D is a transaction database or transaction set and transaction, T , is a subset of items ($T \subseteq I$). Let A be a set composed of items, which is called itemset. Transaction T contains itemset A if and only if $A \subseteq T$. If itemset A contains k items, it is called k -itemset. The number of times itemset A appears in the transaction database D is called the number of supports (σ) for itemset A . The number of occurrences of itemset A in transaction database D as a percentage of the total transactions in D is called the support of itemset A . If the support of an itemset exceeds the minimum support set by the user, it is called a frequent itemset.

The logical implication of association rules can be expressed as $X \Rightarrow Y$, where X is the antecedent of association rules (LHS), Y is the consequent of association rules (RHS) and $X \subset I, Y \subset I, X \cap Y = \emptyset$. If $S\%$ of transactions in transaction database D contain $X \cup Y$, then the support of association rule $X \Rightarrow Y$ is $S\%$. At the same time, if $C\%$ of itemset X contains $X \cup Y$, then the confidence of association rule $X \Rightarrow Y$ is called $C\%$. Thus, support and confidence can be expressed as:

$$Support(X \Rightarrow Y) = P(X \cup Y) = \sigma(X \cup Y)/D \quad (2)$$

Support($X \Rightarrow Y$)—Support of association rule $X \Rightarrow Y$;

$P(X \cup Y)$ —Probability of simultaneous occurrence of X and Y , where $X \cup Y$ should be distinguished from concepts in mathematics;

$\sigma(X \cup Y)$ —Number of simultaneous occurrences of X and Y in datasets;

D —Total number of records in dataset D .

$$\text{Confidence}(X \Rightarrow Y) = P(Y|X) = \sigma(X \cup Y) / \sigma(X) \quad (3)$$

$\text{Confidence}(X \Rightarrow Y)$ —Confidence of association rule $X \Rightarrow Y$;

$P(Y|X)$ —Probability of Y when x appears;

$\sigma(X \cup Y)$ —Probability of simultaneous occurrence of X and Y ;

$\sigma(X)$ —Number of occurrences of X in dataset D .

Association rules mining mainly includes two steps [63]:

- (1) Find all frequent itemsets, that is, all itemsets satisfying minimum support.
- (2) Find strong association rules from frequent itemsets, that is, support and confidence meet the user's threshold.

Apriori algorithm, which is the most common algorithm, is designed based on association rules mining in two steps: the first step is to find all frequent itemsets in the transaction database through iterative loop; the second step is to construct rules that satisfy minimum confidence using frequent itemsets. The core of Apriori algorithm, the implementation process of mining all frequent itemsets, is shown in Figure 4. The basic realization steps are the following:

- (1) Input datasets and user-set minimum support min_sup .
- (2) Scan the dataset, calculate the support of each itemset and generate a set composed of frequent one-itemsets L_1 .
- (3) Perform the step for connecting. In order to form a set composed of frequent k -itemsets, generating a set composed of candidates frequent k -itemsets is a prerequisite. Suppose $m, n \in L_{k-1}$, $m = \{m_1, m_2, \dots, m_{k-2}, m_{k-1}\}$, $n = \{n_1, n_2, \dots, n_{k-2}, n_{k-1}\}$, and when $1 \leq i \leq k-1$, $m_i = n_i$; when $i = k-1$ and $m_i \neq n_i$, then $m \cup n = \{m_1, m_2, \dots, m_{k-2}, m_{k-1}, n_{k-1}\}$ is a candidate frequent k -itemset, which is also an element of C_k .
- (4) Perform the step for pruning. C_k is a superset of L_k , which means some elements of C_k may not be frequent. When C_k is large, it will bring a huge amount of calculation. In this regard, it is a good method to reduce the size of C_k by using the nature of association rules: "superset of non-frequent itemset is still non-frequent itemset". That is to say, when a $k-1$ subset of the candidate frequent k -item set is not an element in L_{k-1} , it shows that the candidate frequent k -item set is also non-frequent and can be removed from C_k .
- (5) Rescan the dataset and calculate the support of each candidate itemsets in C_k .
- (6) Eliminate the itemsets that do not meet the minimum support in C_k to form a set L_k composed of frequent k -itemsets.
- (7) Through iterative loop, steps (3)–(6) above are repeated until the set of new frequent itemsets (non-empty sets) cannot be generated. At this point, Apriori algorithm finds all frequent itemsets satisfying the minimum support.

After all frequent itemsets are mined, they can be directly used to generate strong association rules satisfying the minimum support, min_sup , and the minimum confidence, min_conf :

- (1) For each frequent itemset l , generate all the non-empty subsets of l .
- (2) For each non-empty subset of l , output the rule " $s \Rightarrow (l - s)$ " if $\sigma(l) / \sigma(s) \geq \text{min_conf}$.

However, association rules mining based on traditional Apriori algorithm usually produces more users' uninterested rules or deceptive rules from frequent itemsets under the support-confidence framework, resulting in low efficiency of actual mining. In view of the shortcomings of classical Apriori algorithm, this paper proposes constraint Apriori algorithm to improve it. By adding constraint steps that reflect the actual needs of users in

Apriori algorithm, the generation of useless rules is effectively reduced, and the mining efficiency of association rules is improved.

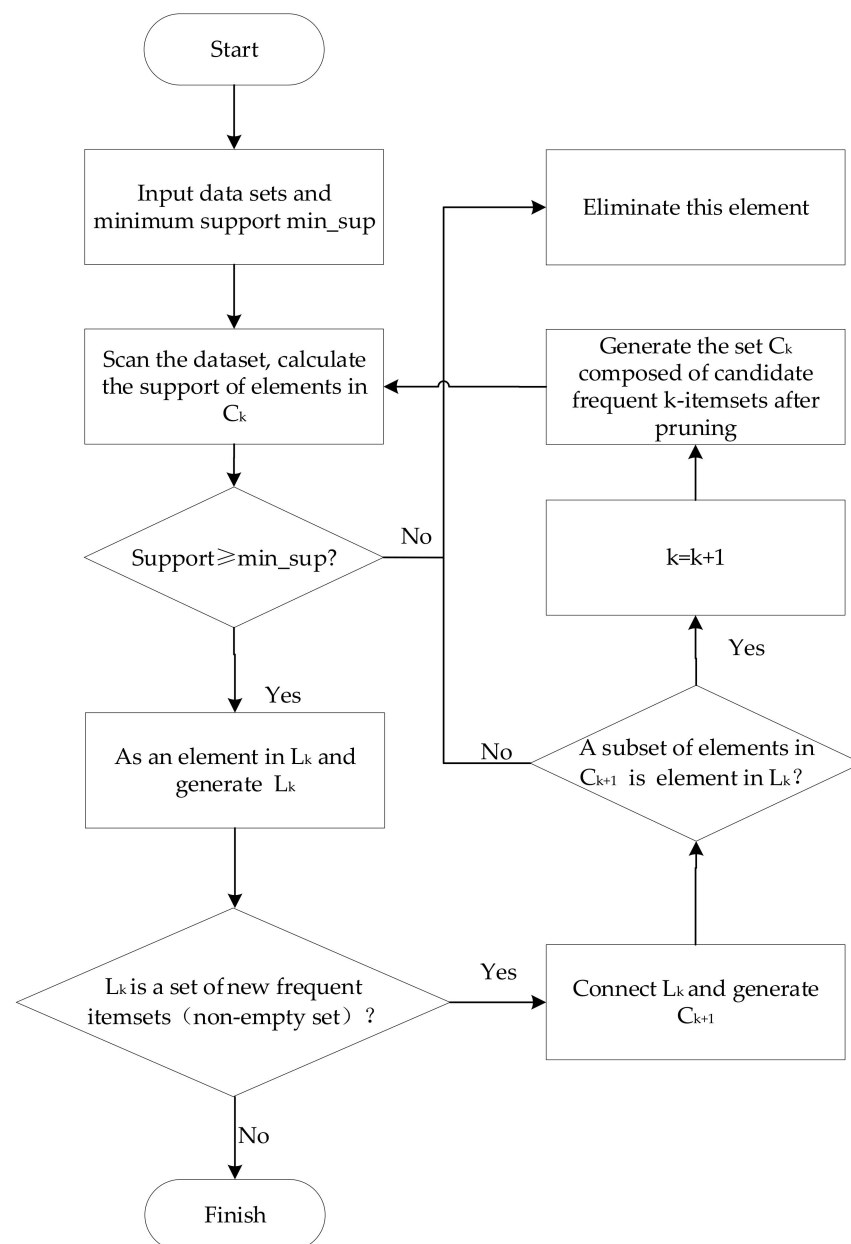


Figure 4. Implementation process of mining all frequent itemsets using Apriori algorithm.

In the current research [64,65], there are five types of common constraint relations:

- (1) Interest degrees constraints: reflect users' interest in rules, such as basic support and confidence;
- (2) Rule constraints: specify the form of mining rules and emphasize rule templates, including the number of assertions, property relationships, property values that appear in the antecedent and consequent of association rules;
- (3) Knowledge type constraints: constrain the type of mining knowledge, such as association rules;
- (4) Data constraints: limit the mined dataset;
- (5) Dimension or layer constraints: describe data dimension or abstract level for mining rules.

The constraint relationship can be divided into objective interest constraint and subjective interest constraint from the subjective and objective perspectives. The objective

interest constraint will help to eliminate some inconsistent rules, and the most commonly used objective interest measure is lift. Lift reflects the influence of rule antecedents on rule consequents, which is expressed as:

$$lift(X \Rightarrow Y) = P(Y|X)/P(Y) = \sigma(X \cup Y)/\sigma(X) \times \sigma(Y) \quad (4)$$

$lift(X \Rightarrow Y)$ —Lift of association rule $X \Rightarrow Y$;

$P(Y|X)$ —Probability of Y when x appears;

$P(Y)$ —Expected confidence of Y , represents the probability of Y in dataset D ;

$\sigma(X \cup Y)$ —Probability of simultaneous occurrence of X and Y ;

$\sigma(X)$ —Number of occurrences of X in dataset D ;

$\sigma(Y)$ —Number of occurrences of Y in dataset D .

The $P(Y)$ in the formula is called the expected confidence of the rule consequent Y , which describes the support of the rule consequent Y itself without the rule antecedent X . The larger the lifting degree is, the greater the influence of rule antecedent X on rule consequent Y is. In general, the lifting degree of effective association rules should be greater than 1, and only when the confidence degree of association rules is greater than the expected confidence degree can the emergence of X promote the emergence of Y , and the also shows that there is a certain degree of correlation between them. If the degree of action is not greater than 1, this association rule is usually meaningless.

Subjective interest reflects researchers' subjective attention to specific association rules for some reasons, which is more closely related to researchers themselves. In the study of this paper, since more attention is paid to the environment pollutants of the construction site, it is more desirable to take the indicators reflecting the pollution as the consequent of the association rules to a subjective extent. In order to achieve this goal, rules constraints can be added to the Apriori algorithm, rule templates are introduced, and the item sets of rule consequents are limited so as to reduce the mining of invalid rules.

Considering all the constraints, this paper adds the lifting indicator and rule template which reflect the objective interest and subjective interest in Apriori algorithm to form the constraint Apriori algorithm and then mine the subsequent association rules. This paper mainly uses R language and R Studio software to realize rule mining based on constraint Apriori algorithm. In R Studio software, the R package rules related to association rules mining are called, and then the dataset is imported. The parameters and conditions required for the application of constrained Apriori algorithm are set. Finally, the program is run to mine and visualize the rules.

3.3. Random Forest

Using association rule mining alone cannot build a readable and verifiable model, and the strong association rules obtained from association rule mining can be further verified utilizing the random forest algorithm to improve the accuracy rate. Random forest is a supervised machine learning algorithm consisting of a decision tree and a bagging framework, where each tree is independent of the other [66]. "Random" in random forest means that samples are randomly selected from the original data to form a subset of data, with the row variables and column observations being randomly drawn each time. "Forest" in random forest is a collection of decision trees in the form of a single decision tree, each of which varies depending on the features used in its construction. While decision trees individually are less effective in solving classification problems, the output of the random forest algorithm is superior.

The random forest construction process is roughly as follows: Firstly, the sample set is randomly drawn from the dataset as the training set according to the Bagging algorithm with release, ensuring that the samples in each training set are not identical. Secondly, the corresponding decision trees are generated according to each training set, and some of the features in the corresponding features of the samples are randomly selected to train the decision trees. Finally, the voting results of each decision tree are counted, and the plurality

or average is taken as the voting result to get the final classification. The introduction of two randomnesses, random features and random data, is crucial to the classification performance of random forests which means the random forest does not easily fall into overfitting and has good noise immunity. To validate the accuracy of the strong association rules obtained from association rule mining, the input variables of the random forest model constructed in this paper are the antecedents of each strong association rule, the output variables are the consequents of each strong association rule and the accuracy threshold of the prediction results is set to filter the rules.

3.4. Jena Reasoning Rules

The typical ontology reasoning machine systems are Jena, Jess, Pellet, Racer and FACT++. In these reasoning machine systems, Jena reasoning machine is more comprehensive and supports multiple languages. It integrates various operations on ontology and can realize ontology parsing, storage, query and other functions. It can also reason based on user-defined rules. Therefore, this paper uses the Jena reasoning machine.

Jena reasoning machine itself has some general rules based on ontology characteristics. However, these generic rules cannot describe the structured and unstructured relationships for a specific domain or in a specific task, so custom construction of the relevant rules using a machine-understandable language is also needed. The rules applied in reasoning can be described by traditional grammar as IF <condition> THEN <conclusion>, that is, when the premise conditions of the rules are met, the rules are executed and the reasoning results are generated. Yet this syntax structure cannot be combined with ontology description language, so it is necessary to transform the expression of rules. At present, the common rule description language compatible with ontology description language is the language structure used by SWRL [67] and Jena custom rules. This paper uses the rule language structure based on Jena to write and express rules.

In the Jena reasoning machine, the reasoning rule is defined as the rule object in the Java system, which is composed of three parts—premise, condition and conclusion—and satisfies the triple pattern. The basic expression is:

$$[rule_{name} : (x P_1 y)(y P_2 z) \rightarrow (x P_3 z)] \quad (5)$$

Among them, $rule_{name}$ is the rule name; $(x P_1 y)(y P_2 z)$ is the subject of the rule; $(x P_3 z)$ is the head of the rule; P_1, P_2, P_3 describe the relationship between x and y , y and z , and x and z , respectively; and the relationship described by P_3 is obtained by reasoning, which means that the implicit relationship between x and y can be obtained by reasoning. Therefore, based on this rule framework, this section will refer to the relevant national norms, the association rules excavated above and the relevant elements in the ontology model to represent the rules in the field of environmental pollutants monitoring in a construction site and then construct the rule database. Some important rules are shown in Table 4. Among them, use En as the abbreviation of <http://www.owl-ontologies.com/Environment.owl> (accessed on 15 December 2020); considering the program's recognition exception for symbols, "PM2.5" is replaced with "PM2".

Table 4. Examples of Jena reasoning rules.

Description of Rules	Representation of Rules
When pH is in the range of 6–9, the pH emission standard grade is 1	[rule_4:(?x rdf: type En# pH) (?x En# has pH Value ?y) greater Than (?y 6) less Than (?y 9) -> (?x En# has Monitoring Level En# Level 1)]
When the NH3—N concentration is less than 15, the NH3—N emission standard level is 1	[rule_6:(?x rdf: type En# NH3-N) (?x En# has Concentration of NH3-N ?y) less Than (?y 15) -> (?x En# has Monitoring Level En# Level 1)]

3.5. Identification Mechanism

The identification mechanism of environmental pollutants in construction site monitoring is mainly based on the matching of ontology individual data and association rules to obtain reasoning results. The main design ideas are as follows:

- (1) The environmental monitoring data of the construction site are added as an individual to the built ontology model in order to update the ontology library;
- (2) The environmental monitoring data of the construction site are backed up in the database, the association rules are analyzed after the pretreatment of the monitoring data in the database, and the effective association rules are obtained and imported into the rule base to achieve real-time updates;
- (3) The updated ontology model and custom rules are loaded into the ontology reasoning machine to perform parsing and reading operations to form a model object with reasoning mechanism;
- (4) The reasoning query results, mainly including judging whether the value of the monitoring indicator exceeds the threshold and judging whether the monitoring indicator level is reasonable, are returned based on reasoning rules. If the monitoring value is within the threshold range and the indicator level is reasonable, the good environmental information is output; if the monitoring value exceeds the threshold and the indicator level is reasonable, the environmental risk information is output; if the indicator level is unreasonable, the reminding information that monitoring may be abnormal is output, and managers are responsible for confirming the monitoring. The specific strategy is explained in Figure 5.

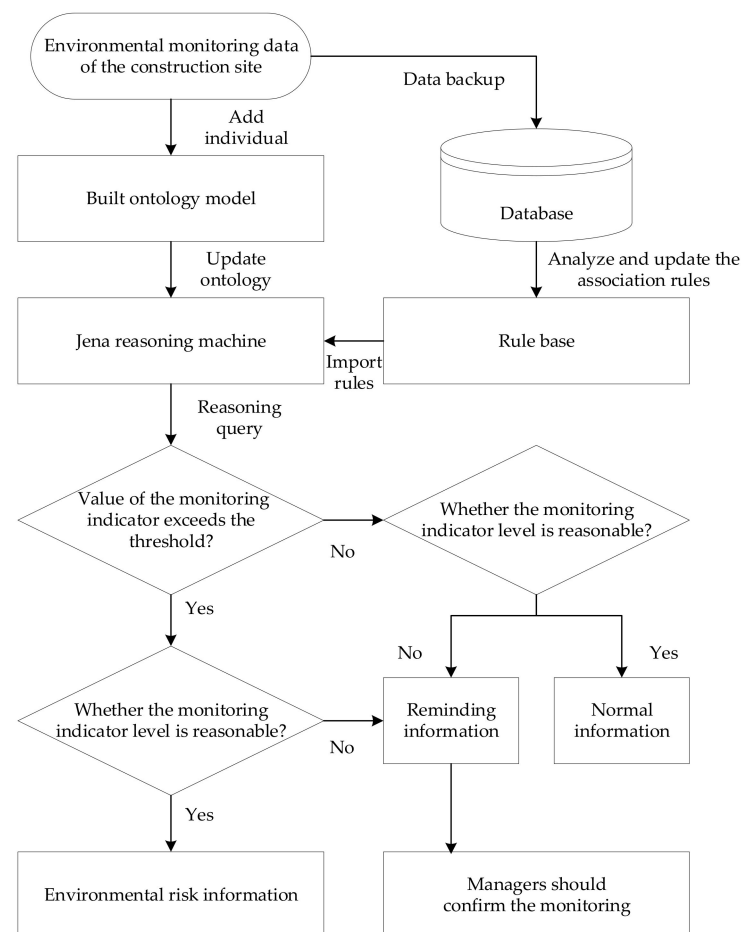


Figure 5. Design ideas of identification mechanism.

4. A Case Study

This paper takes a real estate development project (Plot NO. 2019G83) which covers an area of about 43,000 square meters in Jiangning District of Nanjing City as the research case to demonstrate the detailed process of association rule mining. This site is located in Hushu Street, Jiangning District, Nanjing, east to North Baota Road, south to the planned road, west to Xixu Road and north to Youyi Road. The exact location of the site is illustrated in Figure 6.



Figure 6. The exact location of Plot NO.2019G83.

4.1. Individual Creation in Ontology

The fifth step of the ontology model building is to create an individual which realizes the correspondence between abstract concepts and specific examples. In this case part of the individuals is shown in Table 5.

Table 5. Individuals database in the ontology model.

Types	Individuals
Construction Project	NO. 2019G83 Afford Housing Project
Environment Impact	Air Pollution of NO. 2019G83 Project
Monitoring Item	Air Environment Monitoring of NO. 2019G83 Project
Monitoring Item	Meteorological Environment Monitoring of NO. 2019G83 Project
Monitoring Indicator	Concentration of TSP in Construction Site of NO. 2019G83 Project
Monitoring Indicator	Atmospheric Pressure in Construction Site of NO. 2019G83 Project
Monitoring Site	Construction Site of NO. 2019G83 Project

4.2. Association Rule Mining and Random Forest

(1) Definition of problem.

The environmental pollution caused by construction projects is more serious, especially the dust pollution. Therefore, this paper takes the dust pollution in the construction site as an example to study the correlation between seven monitoring indicators such as temperature, humidity, atmospheric pressure, wind speed, PM2.5, PM10 and TSP and to excavate association rules.

(2) Collection of data.

Data analysis is obtained by investigating the transportation management center of the intelligent site supervision platform in Nanjing. A total of three pieces of environmental monitoring equipment are arranged at two entrances and exits of the site and at the temporary yard of field materials for real-time monitoring of TSP, PM10, PM2.5, temperature, humidity, atmospheric pressure, wind speed, wind direction and noise. Figure 7 demonstrates the position of the three pieces of environmental monitoring equipment in the construction site.



Figure 7. Position of the environmental monitoring equipment.

This paper derives a total of 59,239 monitoring data from Nanjing Smart Site Supervision Platform for 15 consecutive days from 0:00 on 1 March 2021. Some data are shown in Table 6.

Table 6. Raw data exported by the platform.

Collection Time	Temperature (°C)	Humidity (%)	Atmospheric Pressure (kPa)	Wind Speed (m/s)	PM2.5 ($\mu\text{g}/\text{m}^3$)	PM10 ($\mu\text{g}/\text{m}^3$)	TSP ($\mu\text{g}/\text{m}^3$)
1 March 2021 0:00	13.5	85.9	101.69	0	26	40	46
1 March 2021 0:01	13.4	85.5	101.69	0	27	40	47
1 March 2021 0:02	13.4	85.5	101.69	0	25	38	44
1 March 2021 0:03	13.4	85.4	101.69	0	25	37	44
1 March 2021 0:04	13.4	85.6	101.69	0	25	37	43
1 March 2021 0:05	13.4	85.6	101.69	0	26	39	45
1 March 2021 0:06	13.4	85.6	101.68	0	25	37	44
1 March 2021 0:07	13.4	85.7	101.67	0.1	25	38	44
1 March 2021 0:08	13.4	85.2	101.67	0.3	25	38	44
1 March 2021 0:09	13.3	85.4	101.66	0.3	27	41	47
1 March 2021 0:10	13.3	85.4	101.66	0.3	26	40	46

(3) Preprocess of data.

The directly collected data may have some problems such as missing or repeated data or high redundancy or inconsistent data format, so it is crucial to preprocess the data,

including data cleaning, data integration, data reduction, data transformation and other processing methods.

The amount of data obtained in this case is large, and there are some missing or repeated data. Moreover, the application of Apriori algorithm requires the data type to be discrete, so preprocessing the data before mining association rules is completely required, mainly including the preliminary collation and discretization of data.

(i) Data interpretation.

The data collected in this paper are the data gathered at the construction site at a time interval of one minute. The relevant specifications and standards related to TSP, PM10 and PM2.5 usually take the data of every half hour or one hour or one day as the evaluation level, so this paper intends to analyze the monitoring data of each hour. At the same time, in consideration of the small number of missing or repeated data, the influence of these data is ignored. The average value of the actual statistical data per hour is calculated, and the two-digit decimal is retained. Finally, 1015 data are obtained. Some results of data collation are shown in Table 7.

Table 7. Data collation results.

Collection Time	Temperature (°C)	Humidity (%)	Atmospheric Pressure (kPa)	Wind Speed (m/s)	PM2.5 (µg/m ³)	PM10 (µg/m ³)	TSP (µg/m ³)
1 March 2021 1:00	13.17	86.35	101.66	0.04	26.08	39.36	45.81
1 March 2021 2:00	13.18	86.73	101.68	0.03	26.77	40.47	47.02
1 March 2021 3:00	13.19	86.76	101.68	0.01	29.61	45.29	51.63
1 March 2021 4:00	13.07	87.47	101.66	0.02	29.42	44.86	51.27
1 March 2021 5:00	13.12	87.19	101.64	0.02	30.51	46.53	52.85
1 March 2021 6:00	12.89	88.09	101.68	0.32	28.25	42.68	48.93

(ii) Data discretization.

Data discretization is to replace the original values of numerical attributes with interval labels or concept labels and to organize them into higher-level concepts. In this paper, referring to relevant standards, regulations and the literature, the temperature, relative humidity, atmospheric pressure, wind speed, PM2.5, PM10, TSP and other indicators are discretized. The specific classification is shown in Tables 8 and 9.

In the discretization process, the sorted data are discretized and represented in the form of “English alphabet corresponding to the indicator_level”, as shown in Table 10. Taking ‘T_1’ as an example, it indicates that the temperature in the hour is level 1, ranging from 5 to 9.9 °C.

Table 8. Classification of meteorological parameters.

Level	Meteorological Parameter			
	Temperature (°C)	Relative Humidity (%)	Atmospheric Pressure (kPa)	Wind Speed (m/s)
1	0–4.9	0–20	101.0–101.5	0.0–0.2
2	5–9.9	21–40	101.6–102.0	0.3–1.5
3	10–14.9	41–60	102.1–102.5	1.6–3.3
4	15–19.9	61–80	102.6–103.0	3.4–5.4
5	20–24.9	81–100	103.1–103.5	5.5–7.9
6	25–29.9	/	103.6–104.0	8.0–10.7

Table 9. Classification of particulate matter.

Level of Particulate Matter	PM2.5 Concentration ($\mu\text{g}/\text{m}^3$)	PM10 Concentration ($\mu\text{g}/\text{m}^3$)	TSP Concentration ($\mu\text{g}/\text{m}^3$)
1	0.0–12.0	0–50	0–150
2	12.1–35.0	51–150	151–300
3	35.1–55.0	151–250	301–400
4	55.1–150.0	251–350	401–500
5	150.1–250.0	351–420	501–600
6	>250.0	>420	>600

Table 10. Discrete level representation of monitoring indicator.

Monitoring Indicator	Specific Representation of Level
Temperature	T_1, T_2, T_3, T_4, T_5, T_6
Relative humidity	R_1, R_2, R_3, R_4, R_5
Atmospheric pressure	A_1, A_2, A_3, A_4, A_5, A_6
Wind speed	W_1, W_2, W_3, W_4, W_5, W_6
PM2.5	PM2.5_1, PM2.5_2, PM2.5_3, PM2.5_4, PM2.5_5, PM2.5_6
PM10	PM10_1, PM10_2, PM10_3, PM10_4, PM10_5, PM10_6
TSP	TSP_1, TSP_2, TSP_3, TSP_4, TSP_5, TSP_6

(4) Association rule mining.

The constraint Apriori algorithm and the R language implementation steps are used to mine the association rules; the preprocessed dataset and the minimum support, confidence and lift are input. Meanwhile, the consequent of the association rules is constrained to the corresponding levels of PM2.5, PM10 and TSP, and then the strong association rules are output.

Setting the appropriate minimum support and confidence is also the key to output effective strong association rules. If minimum support and confidence are set too low, a large number of invalid rules are mined; if the minimum support and confidence are set too high, some effective knowledge will be filtered out, and relatively comprehensive rules cannot be obtained. At present, in the actual data mining process, the minimum support and confidence are usually set according to industry characteristics and expert experience, or different values are set for the minimum support and confidence, and multiple tests are carried out to determine the appropriate support and confidence according to the test results.

(5) Strong association rules generation

Referring to the opinions of many experts and combining with the experimental analysis, this paper sets the minimum support and minimum confidence as 0.2 and 0.6, respectively, and a total of 24 association rules are excavated, as shown in Table 11.

Table 11. Strong association rules.

Order Number	Strong Association Rules	Support	Confidence	Lift
1	T_3, W_1 $\Rightarrow$ PM2.5_4	0.21	0.69	1.35
2	W_1 $\Rightarrow$ PM2.5_4	0.36	0.64	1.25
3	PM2.5_2 $\Rightarrow$ TSP_1	0.32	1.00	1.16
4	TSP_1, A_3 $\Rightarrow$ PM10_2	0.26	0.78	1.13

Table 11. Cont.

Order Number	Strong Association Rules	Support	Confidence	Lift
5	A_4 $\Rightarrow$ PM10_2	0.28	0.77	1.11
6	PM2.5_4, TSP_1 $\Rightarrow$ PM10_2	xd	0.77	1.11
7	TSP_1, W_1 $\Rightarrow$ PM10_2	0.34	0.76	1.10
8	TSP_1, R_5 $\Rightarrow$ PM10_2	0.26	0.76	1.09
9	W_2 $\Rightarrow$ TSP_1	0.39	0.94	1.09
10	PM10_2, A_3 $\Rightarrow$ TSP_1	0.26	0.94	1.09
11	T_2 $\Rightarrow$ PM10_2	0.26	0.74	1.07
12	A_3 $\Rightarrow$ PM10_2	0.27	0.73	1.06
13	T_2 $\Rightarrow$ TSP_1	0.32	0.91	1.06
14	PM10_2 $\Rightarrow$ TSP_1	0.63	0.91	1.06
15	TSP_1 $\Rightarrow$ PM10_2	0.63	0.73	1.06
16	R_5 $\Rightarrow$ PM10_2	0.28	0.72	1.04
17	A_4 $\Rightarrow$ TSP_1	0.32	0.88	1.02
18	PM10_2, T_3 $\Rightarrow$ TSP_1	0.29	0.88	1.02
19	TSP_1, R_4 $\Rightarrow$ PM10_2	0.27	0.70	1.02
20	A_3 $\Rightarrow$ TSP_1	0.33	0.88	1.02
21	PM10_2, W_1 $\Rightarrow$ TSP_1	0.34	0.88	1.02
22	TSP_1, T_3 $\Rightarrow$ PM10_2	0.29	0.70	1.01
23	TSP_1, W_2 $\Rightarrow$ PM10_2	0.27	0.70	1.01
24	R_5 $\Rightarrow$ TSP_1	0.34	0.87	1.01

(6) Random forest

In this paper, the sklearn machine learning package in Python software is adopted to implement the random forest classification algorithm, dividing the original data into training set, validation set and test set in the ratio of 6:2:2. Take the validation of the first strong association rule T_3, W_1 $\Rightarrow$ PM2.5_4 as an example. When training the model, the six features other than PM2.5, i.e., TSP, PM10, temperature, humidity, atmospheric pressure, wind speed, are regarded as input variables, and the rank of PM2.5 is regarded as the target value. The parameters of the algorithm are adjusted several times according to the performance of the trained model on the validation set, such as the number of decision trees in the forest (*n_estimators*), the minimum sample leaf size (*min_sample_leaf*), the maximum depth of the decision tree (*max_depth*), etc. The final accuracy of the model obtained is 94%, and the out-of-bag data validation score is 93%. Finally, taking 20% of the original data as the test set, samples with T rank 3 and W rank 1 were selected as input variables to predict the PM2.5 rank, and the predicted results were compared with the actual results to calculate the accuracy. The threshold is set to 80%, and if the prediction accuracy is higher than the threshold, the rule will be kept; otherwise it will be removed. The 24 strong association rules were verified sequentially according to the above process, and the accuracy rates were all above the threshold values.

The top 5 rules are explained in this paper, ranked by their degree of lift, as shown in Table 12. Through comparative analysis, it is found that the excavated rules are consistent with the actual situation and can be used as effective rules to guide the dust monitoring and control of the construction site.

Table 12. Interpretation of strong association rules.

Strong Association Rules	Meaning
$T_3, W_1 \Rightarrow PM2.5_4$	When the temperature is 20–24.9 °C and the wind speed is 0.0–0.2 m/s, the concentration of PM2.5 is 55.1–150.0 $\mu\text{g}/\text{m}^3$, and the probability of occurrence is 69%.
$W_1 \Rightarrow PM2.5_4$	When the wind speed is 0.0–0.2 m/s, the concentration of PM2.5 is 55.1–150.0 $\mu\text{g}/\text{m}^3$, and the probability of occurrence is 64%.
$PM2.5_2 \Rightarrow TSP_1$	When the concentration of PM2.5 is 12.1–35.0 $\mu\text{g}/\text{m}^3$, the concentration of TSP is 0–150 $\mu\text{g}/\text{m}^3$, and the probability of occurrence is 100%.
$TSP_1, A_3 \Rightarrow PM10_2$	When the concentration of TSP is 0–150 $\mu\text{g}/\text{m}^3$ and the atmospheric pressure is 102.1–102.5 kPa, the concentration of PM10 is 51–150 $\mu\text{g}/\text{m}^3$, and the probability of occurrence is 78%.
$A_4 \Rightarrow PM10_2$	When the atmospheric pressure is 102.6–103.0 kPa, the concentration of PM10 is 51–150 $\mu\text{g}/\text{m}^3$, and the probability of occurrence is 77%.

Summarizing the 24 rules obtained from the analysis, some conclusions can be easily inferred:

- (1) There is a strong correlation between three dust indicators, including PM2.5, PM10 and TSP, and relevant environmental elements, including temperature, relative humidity, atmospheric pressure and wind speed. When the temperature is level 2–3, relative humidity is level 4–5, atmospheric pressure is level 3–4 and wind speed is level 1–2, the high probability of PM2.5, PM10 and TSP concentrations is level 4, level 2 and level 1.
- (2) PM2.5, PM10 and TSP are related to each other. The fourth level concentration of PM2.5, the second level concentration of PM10 and the first level concentration of TSP are corresponding to each other, and the concentration is controlled in 0–150, which meets the construction dust emission standards.

4.3. Reasoning Implementation

Making use of the constructed ontology model and rule database, this section will use Eclipse platform to realize identification of environmental pollutants based on rule reasoning. The specific tools and software are shown in Table 13.

Table 13. Reasoning tools and software.

Name of Tools and Software	Function
Eclipse IDE for Java Developers	Java-based extensible development platform
Jena—2.6.4	Application development tools in Semantic Web, including related jar packages
MySQL—8.0	Persistent storage of ontology model
MySQL—Connector—Java—8.0	Driver package for connecting MySQL with JDBC
Navicat Premium—15	Database management tool

- (1) Persistent storage of ontology

Although Jena reasoning machine can store the knowledge described by OWL language, its internal memory is small, and the storage effect of data is poor. Therefore, it is generally considered to use relational database MySQL for persistent storage of an ontology model.

(2) Read and call of ontology model

After the ontology model is persistently stored in the database, in order to realize the subsequent reasoning operation of the ontology model, it is also necessary to read the ontology model from the database into OntModel. The reading method is roughly similar to the storage method. When the ontology model is read successfully, the class, property and individual information of the ontology model can be viewed in Eclipse.

(3) Rule reasoning and query

After reading the ontology model, the next steps are to import the rules needed for reasoning, create a general reasoning machine and reasoning model InfModel, write query statements, finally execute reasoning and query. Taking PM10 overrun identification as examples, the program code for the above reasoning process is as follows.

First, call the PM10 identification rule, that is, when the concentration of PM10 exceeds 150, the identification of environmental pollution is expressed in Eclipse as follows: String rule1="[(?x rdf: type http://www.owl-ontologies.com/Environment.owl#PM10)(?x http://www.owl-ontologies.com/Environment.owl#has Concentration of PM10 ?y)greater Than(?y,150)->(?x http://www.owl-ontologies.com/Environment.owl#has State http://www.owl-ontologies.com/Environment.owl#Air Pollution Identification)]".

Secondly, statements for reasoning query is written: String queryString1 = "PREFIX Environment:<http://www.owl-ontologies.com/Environment.owl#> SELECT ?PM10 ?Monitoring Value ?State of Environment WHERE {?PM10 Environment: has State ?State of Environment ?PM10 Environment: has Concentration of PM10 ?Monitoring Value}";

Finally, the general reasoning machine and reasoning model are created, and the inference is performed. The results are shown in Figure 8, indicating that when the PM10 individual data read from the ontology model are 160, the rule-based reasoning can determine that the construction site is in the risk state of air pollution.

```

180= public static void searchOntol(OntModel model) {
181
182     String rule1="[(?x rdf:type http://www.owl-ontologies.com/Environment.owl#PM10)(?x http://www.owl-ontologies.com/Environment.owl#has_Concentration_of_PM10
183         + "greaterThan(?y,150)"
184         + "->(?x http://www.owl-ontologies.com/Environment.owl#has_State http://www.owl-ontologies.com/Environment.owl#Air_Pollution_Early_Warning)]";
185
186
187     String queryString = "PREFIX Environment:<http://www.owl-ontologies.com/Environment.owl#> "
188         + "SELECT ?PM10 ?Monitoring_Value ?State_of_Environment "
189         + "WHERE {?PM10 Environment:has_State ?State_of_Environment ."
190         + "?PM10 Environment:has_Concentration_of_PM10 ?Monitoring_Value} ";
191
192
193     Reasoner reasoner = new GenericRuleReasoner(Rule.parseRules(rule1));
194
195     InfModel inf = ModelFactory.createInfModel(reasoner, model);
196     Query query = QueryFactory.create(queryString);
197
198     QueryExecution qe = QueryExecutionFactory.create(query, inf);
199     ResultSet results = qe.execSelect();
200
201     ResultSetFormatter.out(System.out, results, query);
202     qe.close();

```

Figure 8. PM10 overrun identification.

5. Discussion

Motivated by the above findings and results, the contribution of this study to the existing literature and environmental pollutants monitoring practices includes the following aspects.

Theoretical implications: This study utilizes the ontology approach to express the knowledge in the field of environmental pollutants monitoring in construction sites in a clear, formalized, shared conceptual model and then constructs ontology libraries and rule bases, which can realize the integration, sharing and reuse of relevant knowledge and solve the problem of information silos. In addition, in terms of algorithm, the paper improves the traditional Apriori algorithm, and the proposed constrained Apriori algorithm integrates the subjective and objective interest degrees, which can greatly improve the efficiency of association rule mining and targetedly mine the credible association rules that are of interest to users. In addition, the results of association rules are confirmed by using the random forest algorithm, which combines correlation analysis under big data and

nonlinear modeling, further improving the credibility of data mining results and enriching the method of environmental pollutants monitoring in construction sites.

Practical implications: The association rule mining method can make full use of the environmental monitoring data in the construction site to obtain strong association rules and tacit knowledge between relevant monitoring indicators and enrich the knowledge content in the domain of environmental pollutants monitoring in construction site. Also, the identification mechanism supported by the rule-based reasoning proposed in the paper is able to achieve the intelligent identification of environmental pollutants by using the ontology model and rule base. In particular, depending on the constantly updated association rules, it can infer the possible monitoring values or levels of certain indicators with the help of other related indicators when there are abnormal values or false alarms so that the basis for the identification of environmental pollutants is no longer singular and the accuracy of monitoring has been greatly improved.

Limitations and future research directions: Despite the above theoretical and practical implications, several limitations exist in the current study. First of all, considering the diversity and complexity of knowledge in the field of construction engineering, the presently structured knowledge base is not well developed, and the ontology model is mainly constructed manually; subsequent research may continue to be carried out in terms of the perfection of the knowledge base and the improvement of the construction method. Another point worth noting is that, in practical application, the results generated through identification mechanism can be taken as a reference value, rather than a completely dependent one, and regarded as an absolute standard. Future studies should integrate expert opinion to limit the number of false alarms.

6. Conclusions

The management of environmental pollutants on construction sites is a significant issue that should be urgently addressed. The methodology proposed in this paper first integrates and shares the knowledge in the field of environmental pollutants monitoring at construction sites by building ontology models. Next, the improved Apriori algorithm with added subjective and objective constraints is used for association rule mining among environmental pollutants monitoring indicators. The random forest algorithm is applied to further filter the strong association rules. Finally, the ontology database and rule database are loaded into the Jena reasoning machine for inference, and different signals are issued according to whether different environmental pollutants monitoring indicators exceed the threshold value and the level of the corresponding indicators so as to establish an identification mechanism. By combining ontology-based reasoning and association rule mining, on the one hand, when the data derived by traditional monitoring methods are erroneous or an emergency suddenly occurs, the current risk factors can be inferred and identified based on the monitoring indicators strongly associated with them; on the other hand, the rule base generated by association rule mining is not fixed, but is constantly learning. A continuous update of the rule base is more conducive to improving the accuracy and effectiveness of reasoning.

Author Contributions: Conceptualization, Z.X. and S.P.; methodology, Z.X., H.H. and S.P.; software, H.H. and S.P.; investigation, H.H. and S.P.; writing—original draft preparation, Z.X., H.H. and S.P.; writing—review and editing, Z.X., H.H. and S.P.; visualization, H.H. and S.P.; funding acquisition, Z.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the “National Natural Science Foundation of China” (NSFC-72071043), Natural Science Foundation of Jiangsu Province (BK20201280), MOE (Ministry of Education in China) Project of Humanities and Social Sciences (20YJAZH114), Jiangsu Provincial Construction System Technology Program (2021ZD25).

Data Availability Statement: Data are contained within the article. Additional supporting data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dong, X.; Wu, Y.; Chen, X.; Li, H.; Cao, B.; Zhang, X.; Yan, X.; Li, Z.; Long, Y.; Li, X. Effect of thermal, acoustic, and lighting environment in underground space on human comfort and work efficiency: A review. *Sci. Total. Environ.* **2021**, *786*, 147537. [\[CrossRef\]](#)
2. Matsumoto, Y.; Kubimatsu, S. Evaluation of human perception thresholds of transient vibrations for the assessment of building vibration. *Appl. Acoust.* **2022**, *197*, 108906. [\[CrossRef\]](#)
3. Dräger, P.; Letmathe, P. Value losses and environmental impacts in the construction industry—Tradeoffs or correlates? *J. Clean. Prod.* **2022**, *336*, 130435. [\[CrossRef\]](#)
4. Hong, J.; Hong, T.; Kang, H.; Lee, M. A Framework for Reducing Dust emissions and energy consumption on construction sites. In Proceedings of the 10th International Conference on Applied Energy (ICAE), Västerås, Sweden, 12–15 August 2019; pp. 5092–5096.
5. Hong, J.Y.; Lam, B.; Ong, Z.-T. A multidimensional assessment of construction machinery noises based on perceptual attributes and psychoacoustic parameters. *Autom. Constr.* **2022**, *140*, 104295. [\[CrossRef\]](#)
6. Jung, S.; Kang, H.; Choi, J.; Hong, T.; Park, H.S.; Lee, D.-E. Quantitative health impact assessment of construction noise exposure on the nearby region for noise barrier optimization. *Build. Environ.* **2020**, *176*, 106869. [\[CrossRef\]](#)
7. Hong, J.; Kang, H.; An, J.; Choi, J.; Hong, T.; Park, H.S.; Lee, D.-E. Towards environmental sustainability in the local community: Future insights for managing the hazardous pollutants at construction sites. *J. Hazard. Mater.* **2020**, *403*, 123804. [\[CrossRef\]](#)
8. Kwon, N.; Song, K.; Lee, H.-S.; Kim, J.; Park, M. Construction Noise Risk Assessment Model Focusing on Construction Equipment. *J. Constr. Eng. Manag.* **2018**, *144*, 04018034. [\[CrossRef\]](#)
9. Cheng, B.; Lu, K.; Li, J.; Chen, H.; Luo, X.; Shafique, M. Comprehensive assessment of embodied environmental impacts of buildings using normalized environmental impact factors. *J. Clean. Prod.* **2021**, *334*, 130083. [\[CrossRef\]](#)
10. Hong, T.; Ji, C.; Park, J.; Leigh, S.-B.; Seo, D.-Y. Prediction of Environmental Costs of Construction Noise and Vibration at the Preconstruction Phase. *J. Manag. Eng.* **2015**, *31*, 04014079. [\[CrossRef\]](#)
11. Khamraev, K.; Cheriyan, D.; Choi, J.-H. A review on health risk assessment of PM in the construction industry—Current situation and future directions. *Sci. Total Environ.* **2020**, *758*, 143716. [\[CrossRef\]](#)
12. Azarmi, F.; Kumar, P.; Marsh, D.; Marsh, D.; Fuller, G. Assessment of the long-term impacts of PM10 and PM2.5 particles from construction works on surrounding areas. *Environ. Sci.-Process. Impacts* **2016**, *18*, 208–221. [\[CrossRef\]](#)
13. Yan, H.; Ding, G.; Feng, K.; Zhang, L.; Li, H.; Wang, Y.; Wu, T. Systematic evaluation framework and empirical study of the impacts of building construction dust on the surrounding environment. *J. Clean. Prod.* **2020**, *275*, 122767. [\[CrossRef\]](#)
14. Yan, H.; Ding, G.L.; Li, H.Y. Field Evaluation of the Dust Impacts from Construction Sites on Surrounding Areas: A City Case Study in China. *Sustainability* **2019**, *11*, 1906. [\[CrossRef\]](#)
15. Guo, P.; Tian, W.; Li, H. Dynamic health risk assessment model for construction dust hazards in the reuse of industrial buildings. *Build. Environ.* **2022**, *210*, 108736. [\[CrossRef\]](#)
16. Tao, X.; Mao, C.; Xie, F.; Liu, G.; Xu, P. Greenhouse gas emission monitoring system for manufacturing prefabricated components. *Autom. Constr.* **2018**, *93*, 361–374. [\[CrossRef\]](#)
17. Yang, C.-T.; Chen, H.-W.; Chang, E.-J.; Kristiani, E.; Nguyen KL, P.; Chang, J.-S. Current advances and future challenges of AIoT applications in particulate matters (PM) monitoring and contro. *J. Hazard. Mater.* **2021**, *419*, 126442. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Guo, X.H.; Wang, Y.F.; Mei, S.Q. Monitoring and modelling of PM2.5 concentration at subway station construction based on IoT and LSTM algorithm optimization. *J. Clean. Prod.* **2022**, *360*, 132179. [\[CrossRef\]](#)
19. Hong, J.; Kang, H.; Jung, S.; Sung, S.; Hong, T.; Park, H.S.; Lee, D.-E. An empirical analysis of environmental pollutants on building construction sites for determining the real-time monitoring indices. *Build. Environ.* **2019**, *170*, 106636. [\[CrossRef\]](#)
20. Abdeljaber, O.; Avci, O.; Kiranyaz, S.; Gabbouj, M.; Inman, D.J. Real-time vibration-based structural damage detection using one-dimensional convolutional neural networks. *J. Sound Vib.* **2017**, *388*, 154–170. [\[CrossRef\]](#)
21. Yang, Q.N.; Shi, W.M.; Chen, J. Deep convolution neural network-based transfer learning method for civil infrastructure crack detection. *Autom. Constr.* **2020**, *116*, 103199. [\[CrossRef\]](#)
22. Kim, H.; Tae, S.; Zheng, P.; Kang, G.; Lee, H. Development of IoT-based particulate matter monitoring system for construction sites. *Int. J. Environ. Res. Public Health* **2021**, *18*, 11510. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Arajo, I.P.S.; Costa, D.B.; De Moraes, R.J.B. Identification and characterization of particulate matter concentrations at construction jobsites. *Sustainability* **2014**, *6*, 7666–7688. [\[CrossRef\]](#)
24. Guo, P.; Tian, W.; Li, H.; Zhang, G.; Li, J. Global characteristics and trends of research on construction dust: Based on bibliometric and visualized analysis. *Environ. Sci. Pollut. Res.* **2020**, *27*, 37773–37789. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Agrawal, R.; Imieliński, T.; Swami, A. Mining association rules between sets of items in large databases. *ACM SIGMOD Rec.* **1993**, *22*, 207–216. [\[CrossRef\]](#)
26. Trillo Cabello, A.; Martinez-Rojas, M.; Carrillo-Castrillo, J.A.; Rubio-Romero, J.C. Occupational accident analysis according to professionals of different construction phases using association rules. *Saf. Sci.* **2021**, *144*, 105457. [\[CrossRef\]](#)
27. Zhou, Y.; Li, C.; Ding, L.; Sekula, P.; Love, P.E.; Zhou, C. Combining association rules mining with complex networks to monitor coupled risks. *Reliab. Eng. Syst. Saf.* **2019**, *186*, 194–208. [\[CrossRef\]](#)
28. Verama, A.; Khan, S.D.; Maiti, J.; Krishna, O.B. Identifying patterns of safety related incidents in a steel plant using association rule mining of incident investigation reports. *Saf. Sci.* **2014**, *70*, 89–98. [\[CrossRef\]](#)

29. Xu, R.H.; Luo, F. Risk prediction and early warning for air traffic controllers' unsafe acts using association rule mining and random forest. *Saf. Sci.* **2021**, *135*, 105125. [[CrossRef](#)]
30. Speiser, J.L.; Miler, M.E.; Tooze, J. A comparison of random forest variable selection methods for classification prediction modeling. *Expert Syst. Appl.* **2019**, *134*, 93–101. [[CrossRef](#)]
31. Lee, J.Y.; Kim, K.Y. Semantic and Association Rule Mining-based Knowledge Extension for Reusable Medical Equipment Random Forest Rules. *J. Integr. Des. Process Sci.* **2018**, *22*, 55–81. [[CrossRef](#)]
32. Qu, Z.; Wang, F.; Zhang, Y. Thickness prediction of seismic multi-attributes sand based on association rules and random forests. *Bull. Geol. Sci. Technol.* **2021**, *40*, 211–218.
33. Gruber, T.R. A translation approach to portable ontology specifications. *Knowl. Acquis.* **1993**, *5*, 199–220. [[CrossRef](#)]
34. Elhadj, H.B.; Sallabi, F.; Henaien, A.; Chaari, L.; Shuaib, K.; Al Thawadi, M. Do-Care: A dynamic ontology reasoning based healthcare monitoring system. *Future Gener. Comput. Syst.* **2021**, *118*, 417–431. [[CrossRef](#)]
35. Lfe, C.-H.; Wang, Y.-H.; Trappey, A.J.C. Ontology-based reasoning for the intelligent handling of customer complaints. *Comput. Ind. Eng.* **2015**, *84*, 144–155.
36. Wong, M.S.; Mok, E.; Wang, T.; Yong, Z. Development of an integrated Micro-environmental monitoring system for construction sites. *Procedia Environ. Sci.* **2016**, *36*, 207–214. [[CrossRef](#)]
37. Smaoui, N.; Kim, K.; Gnawali, O.; Lee, Y.-J.; Suh, W. Respirable Dust Monitoring in Construction Sites and Visualization in Building Information Modeling Using Real-time Sensor Data. *Sens. Mater.* **2018**, *30*, 1775. [[CrossRef](#)]
38. Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P. From Data Mining to Knowledge Discovery in Databases. *AI Mag.* **1996**, *17*, 37–54.
39. Han, J.; Pei, J. *Mining Frequent Patterns without Candidate Generation*; ACM: New York, NY, USA, 2000.
40. Zaki, M.J.; Parthasarathy, S.; Ogihara, M.; Li, W. *New Algorithms for Fast Discovery of Association Rules*; AAAI Press: Menlo Park, CA, USA, 1997.
41. Holland John, H. Genetic algorithms and the optimal allocation of trials. *Siam J. Comput.* **1973**, *2*, 88–105. [[CrossRef](#)]
42. Minaei-Bidgoli, B.; Barmaki, R.; Nasiri, M. Mining numerical association rules via multi-objective genetic algorithms. *Inf. Sci.* **2013**, *233*, 15–24. [[CrossRef](#)]
43. Dorigo, M.; Member, S. Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman. *IEEE Trans. Evol. Comput.* **1996**, *1*, 53–66. [[CrossRef](#)]
44. Kuo, R.J.; Chao, C.M.; Chiu, Y.T. Application of particle swarm optimization to association rule mining. *Appl. Soft Comput.* **2011**, *11*, 326–336. [[CrossRef](#)]
45. Sun, C.; Zhang, R.; Sharples, S.; Han, Y.; Zhang, H. Thermal comfort, occupant control behaviour and performance gap—A study of office buildings in north-east China using data mining. *Build. Environ.* **2019**, *149*, 305–321. [[CrossRef](#)]
46. Tao, G.W.; Feng, J.C.; Feng, H.B. Reducing Construction Dust Pollution by Planning Construction Site Layout. *Buildings* **2022**, *12*, 531. [[CrossRef](#)]
47. Borst, W.N. *Construction of Engineering Ontologies for Knowledge Sharing and Reuse*; Universiteit Twente: Twente, The Netherlands, 1997.
48. Studer, R.; Benjamins, V.R.; Fensel, D. Knowledge Engineering: Principles and Methods. *Data Knowl. Eng.* **2008**, *25*, 161–197. [[CrossRef](#)]
49. Feilmayr, C.; Woss, W. An analysis of ontologies and their success factors for application to business. *Data Knowl. Eng.* **2016**, *101*, 1–23. [[CrossRef](#)]
50. El Kharbili, M.; Stolarski, P. Building-up a reference generic regulation ontology: A bottom-up approach. In Proceedings of the 12th International Conference on Business Information Systems, Poznan, Poland, 27–29 April 2009.
51. Mesaric, E.J.; Dukic, B. An approach to creating domain ontologies for higher education in economics. In Proceedings of the 29th International Conference on Information Technology Interfaces, Cavtat, Croatia, 25–28 June 2007.
52. Savonnet, M.; Leclercq, E.; Naubourg, P. eClims: An Extensible and Dynamic Integration Framework for Biomedical Information Systems. *IEEE J. Biomed. Health Inform.* **2015**, *20*, 1640–1649. [[CrossRef](#)]
53. Li, Y.; Ouyang, S.; Zhang, Y. Combining deep learning and ontology reasoning for remote sensing image semantic segmentation. *Knowl.-Based Syst.* **2022**, *243*, 108469. [[CrossRef](#)]
54. Saraiva, R.; Perkusich, M.; Silva, L.; Almeida, H.; Siebra, C.; Perkusich, A. Early diagnosis of gastrointestinal cancer by using case-based and rule-based reasoning. *Expert Syst. Appl.* **2016**, *61*, 192–202. [[CrossRef](#)]
55. Ding, L.Y.; Zhong, B.T.; Wu, S.; Luo, H.B. Construction risk knowledge management in BIM using ontology and semantic web technology. *Saf. Sci.* **2016**, *87*, 202–213. [[CrossRef](#)]
56. Tserng, H.P.; Yin, S.Y.L.; Dzeng, R.J.; Wou, B.; Tsai, M.D.; Chen, W.Y. A study of ontology-based risk management framework of construction projects through project life cycle. *Autom. Constr.* **2009**, *18*, 994–1008. [[CrossRef](#)]
57. Moradi, H.; Sebt, M.H.; Shakeri, I.E. Toward improving the quality compliance checking of urban private constructions in Iran: An ontological approach. *Sustain. Cities Soc.* **2018**, *38*, 137–144. [[CrossRef](#)]
58. Uschold, M.; Gruninger, M. Ontologies: Principles, methods and applications. *Knowl. Eng. Rev.* **1996**, *11*, 93–136. [[CrossRef](#)]
59. Liu, J.E.; Zheng, B.J.; Luo, L.M. Ontology representation and mapping of common fuzzy knowledge. *Neurocomputing* **2016**, *215*, 184–195. [[CrossRef](#)]
60. Lopez, M.F.; Gomez-Perez, A.; Sierra, J.P. Building a chemical ontology using methontology and the ontology design environment. *IEEE Intell. Syst. Appl.* **1999**, *14*, 37–46. [[CrossRef](#)]

-
61. Starr, R.R.; De Oliveira, J.M.P. Concept maps as the first step in an ontology construction method. *Inf. Syst.* **2013**, *38*, 771–783. [\[CrossRef\]](#)
 62. Agrawal, S.R. Mining generalized association rules. *Future Gener. Comput. Syst.* **1997**, *13*, 161–180.
 63. Witten Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed.; Morgan Kaufmann: Boston, MA, USA, 2011.
 64. Guang-Yuan, L.; Dan-Yang, C.; Jian-Wei, G. Association Rules Mining with Multiple Constraints. *Procedia Eng.* **2011**, *15*, 1678–1683. [\[CrossRef\]](#)
 65. Baralis, E.; Cagliero, L.; Cerquitelli, T.; Garza, P. Generalized association rule mining with constraints. *Inf. Sci.* **2012**, *194*, 68–84. [\[CrossRef\]](#)
 66. Giri, I.S.; Kang, Y.; Macdonald, K.; Tippet, M.; Qiu, Z.; Lathrop, R.G.; Obropta, C.C. Revealing the sources of arsenic in private well water using random forest classification and regression. *Sci. Total Environ.* **2023**, *857*, 159360. [\[CrossRef\]](#)
 67. Abadi, A.; Ben-Azza, H.; Sekkat, S. Improving integrated product design using SWRL rules expression and ontology-based reasoning. *Procedia Comput. Sci.* **2018**, *127*, 416–425. [\[CrossRef\]](#)