PROCEEDINGS OF SCIENCE

Weather and Landslide Data Analysis Using Spatio-Temporal and Categorical Data Mining

Yasuhiro Hayashi[a,*]

[a] *Department of Data Science, Musashino University, 3-3, Ariake 3, Koto-Ward, Tokyo, Japan*

E-mail: yhayashi@musashino-u.ac.jp

This paper proposes a semantic correlation and knowledge mining method that identifies semantic correlation structures between two sets with different granularities and numbers of elements, and performs semantic interpretation and knowledge representation of coarse-grained events. When humans understand an event, we do not refer only to a single element that directly corresponds to the event. Instead, use the time before and after the event, the space in which the event occurred, and related concepts as clues. We then select semantically related subsets from a large amount of surrounding information and integrate them as the background of the event. The proposed method aims to computationally formalize this process of background exploration. The proposed method targets a set A, which contains many fine-grained elements, and a set B, which contains a small number of coarse-grained elements. For each $b_j \in B$, the method constructs a semantically related subset $A_j \subseteq A$ from set A. Each element is represented in a semantic space S composed of time, space, and vocabulary, and the semantic correlation between the aggregated representation of the subset and the representation of b_j is computed. Through this process, the proposed method explicitly describes latent correspondences between data sets with different granularities and representation formats as subset-level correlation structures, and provides a framework for transforming these structures into reusable knowledge representations. Furthermore, this paper applies the proposed method to the disaster prevention domain and demonstrates that semantic correlation structures can be constructed between meteorological observation data and landslide event data. The results confirm that the proposed method is effective for the integrated understanding of heterogeneous data sets and for knowledge extraction.

International Symposium on Grids and Clouds (ISGC2026)
15-20, March, 2026
Academia Sinica, Taipei, Taiwan

*Speaker

© Copyright owned by the author(s) under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0) All rights for text and data mining, AI training, and similar technologies for commercial purposes, are reserved. ISSN 1824-8039 . Published by SISSA Medialab.

PoS(ISGC2026)001

1. Introduction

Identifying the relationships that exist among multiple datasets with heterogeneous characteristics is important for solving real-world problems. For example, in disaster prevention, weather data are continuously observed at high frequency and over wide areas, whereas landslide data are recorded at relatively coarse granularity in terms of occurrence location, occurrence time, and damage status. An asymmetry exists between these two data: one set contains a large number of fine-grained elements, while the other contains a small number of coarse-grained elements. When humans interpret an event, we do not rely on a single element that directly corresponds to the event. Instead, select multiple semantically related elements from surrounding information based on temporal context, spatial context, and related concepts, and integrate them to construct the background of the event.

This paper proposes a semantic correlation and knowledge mining method that identifies semantic correlations between two sets with different characteristics, namely, a fine-grained set A with many elements and a coarse-grained set B with a small number of elements, and performs semantic interpretation and knowledge representation of low-frequency, coarse-grained events. In the disaster domain, set A corresponds to weather observation sequences, and set B corresponds to landslide events. Each landslide event should not be explained by a single observation at a specific time point, but rather by a subset of multiple weather events observed prior to the occurrence. The objective of this study is to formalize the human process of background exploration —namely, selecting temporal, spatial, and conceptual context surrounding an event, aggregating it, and transforming it into an explanation—by representing it as $b_j \to A_j \subseteq A \to \mathrm{Agg}(A_j) \to \mathrm{Corr}(\mathrm{Agg}(A_j), b_j)$, thereby enabling machines to perform semantic interpretation and knowledge extraction in a manner analogous to human cognition. Here, semantic correlation denotes the process of identifying semantic relationships among elements or subsets, and knowledge mining denotes the process of extracting and representing interpretable knowledge from such correlation structures. A semantic space S composed of time, space, and vocabulary is introduced. Each element in sets A and B is mapped onto a representation in this space. Temporal and spatial components function as reference coordinate systems for aligning data with different granularities and representation formats, while the vocabulary component provides symbolic representations that encode domain-specific meanings. This paper introduces a framework in which each $b_j \in B$ is explained not by a single corresponding element, but by a subset $A_j \subseteq A$ constructed from set A. This enables the relationships between sets with different granularities to be represented not at the element level, but at the subset level. The core of this method lies in aligning the temporal, spatial, and vocabulary components of the extracted representations within the semantic space S, and explicitly describing the correspondence and correlation structures between two sets with different granularities and representation formats. In this sense, the proposed method is positioned as a framework for "semantic bridging" that connects heterogeneous data sets.

To address the comparison of asymmetric sets, previous studies on landslide prediction have widely employed rainfall threshold models based on rainfall intensity, duration, and accumulation; physical models considering slope stability and groundwater behavior; statistical models based on historical trends; and machine learning and deep learning models such as Random Forest, SVM, XGBoost, ANN, and DNN [1]. Additionally, studies incorporating soil moisture to improve early

PoS(ISGC2026)001

warning performance and studies applying time-series models such as LSTM and TCN for displacement prediction have also been conducted [1,2]. Although these studies have achieved important results in predicting the probability or risk level of specific events, most assume correspondences based on individual observations or feature vectors. A framework that explicitly constructs subset-level correspondences between two sets with significantly different granularities and element sizes, and represents them as semantic correlation structures, has not been sufficiently explored.

The author et al. have previously proposed a Spatio-Temporal and Categorical Data Mining Method for computing relationships between heterogeneous sets based on spatial, temporal, and categorical dimensions [3]. This study introduced a framework in which elements from two sets are mapped to temporal, spatial, and categorical features, and the strength of relationships is computed using distance measures and normalized integrated scores. It also unified deductive analysis, which performs retrieval based on given hypotheses, and inductive analysis, which discovers hypotheses from data, through the use of context vectors.

This paper further demonstrates the effectiveness and feasibility of the proposed method by implementing it in the disaster prevention domain using meteorological data and landslide data. The proposed method is designed as a higher-level concept that can be implemented and evaluated differently depending on the target domain. Therefore, it is applicable not only to disaster-related data such as weather and landslides, but also to a wide range of domains that require the analysis of relationships between heterogeneous event sets with different granularities, such as business, education, and human behavior understanding.

2. A Semantic Correlation and Knowledge Mining Method

2.1 Component Extraction Functions and Element Representation

The proposed method considers two sets A and B:

$$A = \{a_1, a_2, \ldots, a_n\}, \qquad B = \{b_1, b_2, \ldots, b_m\} \tag{1}$$

Here, set A is a fine-grained set with a large number of elements, whereas set B is a coarse-grained set with a small number of elements. The proposed method adopts a framework in which each element $b_j \in B$ is explained by associating it with a subset $A_j \subseteq A$.

For each element $x \in A \cup B$, the proposed method defines component extraction functions for temporal, spatial, vocabulary, and auxiliary attributes as follows:

$$t : A \cup B \to T \quad \ell \ : A \cup B \to L \quad v : A \cup B \to V \quad u \ : A \cup B \to U \tag{2}$$

Here, $t(x)$ denotes the temporal component, $\ell(x)$ denotes the spatial component, $v(x)$ denotes the vocabulary component, and $u(x)$ denotes the auxiliary attribute component.

The proposed method defines a unified descriptive representation $\mathcal{X}(x)$ as:

$$\mathcal{X}(x) = \big(t(x),\ \ell(x),\ v(x),\ u(x)\big) \tag{3}$$

$\mathcal{X}(x)$ serves as a fundamental representation used for semantic mapping, subset construction, aggregation, and correlation computation. The auxiliary attribute component $u(x)$ represents domain-dependent information and is optionally used in subset construction and correlation computation.

PoS(ISGC2026)001

2.2 Semantic Space S

A semantic space S is introduced to uniformly represent semantic relationships.

$$S = T \times L \times V \tag{4}$$

Each component space is equipped with a distance function:

$$d_T : T \times T \to \mathbb{R}_{\geq 0} \quad d_L \ : L \times L \to \mathbb{R}_{\geq 0} \quad d_V : V \times V \to \mathbb{R}_{\geq 0} \tag{5}$$

The proposed method maps each element $x \in A \cup B$ to a point in the semantic space S using $(t(x), \ell(x), v(x))$. The auxiliary attribute $u(x)$ is treated as an external component and is not included in S. The specific representations of T, L, and V are defined depending on the domain.

2.3 Subset Construction

The proposed method constructs a subset $A_j \subseteq A$ for each $b_j \in B$.

$$\sigma_A : B \times \Theta \to 2^A \tag{6}$$

$$A_j = \sigma_A(b_j; \theta) = \{a \in A \mid C(a, b_j; \theta) = \text{true}\} \tag{7}$$

$$C : A \times B \times \Theta \to \{\text{true}, \text{false}\} \tag{8}$$

The predicate $C(a, b_j; \theta)$ is defined as a conjunction of temporal, spatial, vocabulary, and attribute conditions.

$$\theta = (\theta_t, \theta_\ell, \theta_v, \theta_u) \tag{9}$$

2.4 Aggregation Representation

The proposed method defines mappings to the semantic space:

$$\phi_A : A \to S, \qquad \phi_B : B \to S \tag{10}$$

The aggregation function is defined as:

$$\Gamma_A : 2^A \to \mathbb{R}^d \tag{11}$$

$$r_j^{(A)} = \Gamma_A(A_j) \tag{12}$$

The transformation for B is defined as:

$$\psi_B : S \to \mathbb{R}^d \tag{13}$$

$$r_j^{(B)} = \psi_B(\phi_B(b_j)) \tag{14}$$

Thus, both representations are mapped into a common feature space $\mathbb{R}^d$.

PoS(ISGC2026)001

2.5 Correlation Function

The correlation function is defined as:

$$\Psi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^k \tag{15}$$

$$c_j = \Psi(r_j^{(A)}, r_j^{(B)}) \tag{16}$$

2.6 Knowledge Mining

The correlation structure is defined as:

$$C_B = \left\{(b_j, A_j, c_j, \bar{c}_j) \mid b_j \in B\right\} \tag{17}$$

The knowledge extraction mapping is:

$$\mathcal{M} : C_B \to K \tag{18}$$

$$K = K_{\text{rule}} \cup K_{\text{pattern}} \cup K_{\text{concept}} \cup K_{\text{prototype}} \cup K_{\text{index}} \tag{19}$$

2.7 Retrieval and Prediction

The current subset is defined as:

$$A^* = \sigma_A^*(q^*; \theta) \tag{20}$$

$$r^* = \Gamma_A(A^*) \tag{21}$$

The similarity function is defined as:

$$\text{Sim} : \mathbb{R}^d \times \mathbb{R}^d \to [0, 1] \tag{22}$$

$$s_j = \text{Sim}(r^*, r_j^{(A)}) \tag{23}$$

$$N_k(r^*) = \arg\text{top}_k \left\{\text{Sim}(r^*, r_j^{(A)})\right\} \tag{24}$$

$$P(y \mid r^*) = \frac{\left|\{j \in N_k(r^*) \mid \text{label}(b_j) = y\}\right|}{k} \tag{25}$$

PoS(ISGC2026)001

3. Implementation for Disaster Prevention Domain

This section concretizes the Semantic Correlation and Knowledge Mining method defined in Section 2 by applying it to the disaster prevention domain. Each mapping and function defined in the abstract framework is implemented as actual data processing procedures and database structures. The implementation defines a set A consisting of fine-grained and high-frequency meteorological observation events, and a set B consisting of coarse-grained and low-frequency landslide events. Each $b_j \in B$ is explained by a subset $A_j \subseteq A$ composed of meteorological events observed prior to its occurrence.

The implementation targets precursor analysis of landslide disasters. The implementation defines set A as a collection of meteorological observations recorded at regular intervals, and set B as a collection of recorded landslide events. Each element $a_i \in A$ includes attributes such as observation time, observation location, precipitation, temperature, humidity, wind speed, and optionally soil moisture. Each element $b_j \in B$ includes attributes such as occurrence time, occurrence location, damage scale, damage type, recovery duration, and affected area. The implementation constructs a subset $A_j \subseteq A$ for each $b_j \in B$ based on the subset extraction function defined in Section 2. Thus, each landslide event is explained not by a single observation but by a semantically related subset of meteorological events.

3.1 Semantic Representation

The implementation instantiates the component extraction functions defined in Section 2.1 for the disaster prevention domain. For each meteorological event $a_i \in A$: $t(a_i)$ denotes the observation timestamp, $\ell(a_i)$ denotes the latitude and longitude, $v(a_i)$ denotes vocabulary-based descriptions of weather conditions, $u(a_i)$ denotes numerical attributes such as precipitation, temperature, humidity, wind speed, and soil moisture. For each landslide event $b_j \in B$: $t(b_j)$ denotes the occurrence time, $\ell(b_j)$ denotes the geographic coordinates, $v(b_j)$ denotes landslide type and damage descriptions, $u(b_j)$ denotes damage-related attributes.

Thus, each element is represented as:

$$\mathcal{X}(x) = (t(x), \ell(x), v(x), u(x)) \tag{26}$$

The semantic space is defined as:

$$S = T \times L \times V \tag{27}$$

The implementation aligns heterogeneous data not only numerically but also semantically by using time and space as alignment axes and vocabulary as a semantic mediator.

3.2 Database Schema

The implementation realizes the abstract model as relational database tables in PostgreSQL.
(i) Weather Observation Table

The table `weather_observations` represents set A: $t(a_i) \to$ `observed_at` $\ell(a_i) \to$ latitude, longitude $v(a_i) \to$ `weather_label` $u(a_i) \to$ numerical attributes.
(ii) Landslide Event Table

PoS(ISGC2026)001

The table `landslide_events` represents set B with corresponding mappings.

(iii) Subset Link Table

The table `event_weather_links` explicitly stores the subset A_j.

(iv) Aggregated Weather Profiles

The table `aggregated_weather_profiles` stores:

$$r_j^{(A)} = \Gamma_A(A_j) \tag{28}$$

(v) Landslide Profiles

The table `landslide_profiles` stores:

$$r_j^{(B)} = \psi_B(\phi_B(b_j)) \tag{29}$$

(vi) Correlation Table

The table `semantic_correlations` stores:

$$c_j = \Psi(r_j^{(A)}, r_j^{(B)}), \quad \bar{c}_j \tag{30}$$

(vii) Knowledge Tables

The implementation stores rule, pattern, prototype, and concept knowledge in relational tables.

3.3 Feature Definition

The implementation defines the aggregation:

$$r_j^{(A)} = \Gamma_A(A_j) \tag{31}$$

Features are categorized into statistical features, temporal precursor features, spatial precursor features, and vocabulary-based semantic features. Each feature is defined as the following formulas.

(i) Statistical Features

$$\text{avgPrec}(A_j) = \frac{1}{|A_j|} \sum_{a \in A_j} \text{prec}(a) \tag{32}$$

$$\text{sumPrec}(A_j) = \sum_{a \in A_j} \text{prec}(a) \tag{33}$$

(ii) Temporal Features

$$\text{Prec}_h(A_j) = \sum_{a \in A_j : t(a) \geq t(b_j) - h} \text{prec}(a) \tag{34}$$

(iii) Spatial Features

$$\text{nearPrec}(A_j) = \frac{1}{|A_j|} \sum_{a \in A_j} w(d_L(\ell(a), \ell(b_j))) \, \text{prec}(a) \tag{35}$$

(iv) Vocabulary Features

$$\text{heavyRain}(A_j) = \mathbf{1}(\text{Prec}_{24}(A_j) \geq \theta) \tag{36}$$

PoS(ISGC2026)001

3.4 Subset Construction

The implementation instantiates:

$$A_j = \sigma_A(b_j; \theta) \tag{37}$$

The parameters define temporal, spatial, vocabulary, and attribute constraints.

$$\theta = (\theta_t, \theta_\ell, \theta_v, \theta_u) \tag{38}$$

3.5 Aggregation

The implementation transforms:

$$r_j^{(A)} = \Gamma_A(A_j) \tag{39}$$

by combining statistical, temporal, spatial, and semantic features.

3.6 Vocabulary Mapping

The implementation introduces:

$$v_j^{(A)} = \Phi_V(A_j) \tag{40}$$

to map numerical features into semantic vocabulary space.

3.7 Correlation Computation

The Correlation $c_j = \Psi(r_j^{(A)}, r_j^{(B)})$ is implemented as follows:
(i) Temporal Correlation

$$d_t = |t(b_j) - \bar{t}(A_j)| \tag{41}$$

$$c_j^{(t)} = \exp\left(-\frac{d_t^2}{2\sigma_t^2}\right) \tag{42}$$

(ii) Spatial Correlation

$$d_\ell = 2R \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_j - \bar{\phi}_j}{2}\right) + \cos(\phi_j)\cos(\bar{\phi}_j)\sin^2\left(\frac{\lambda_j - \bar{\lambda}_j}{2}\right)}\right) \tag{43}$$

$$c_j^{(\ell)} = \exp\left(-\frac{d_\ell^2}{2\sigma_\ell^2}\right) \tag{44}$$

(iii) Vocabulary Correlation

$$c_j^{(v)} = \frac{v_j^{(A)} \cdot v_j^{(B)}}{\|v_j^{(A)}\| \|v_j^{(B)}\|} \tag{45}$$

PoS(ISGC2026)001

(iv) Attribute Correlation

$$c_j^{(u)} = \exp\left(-\frac{\|r_j^{(A)} - r_j^{(B)}\|^2}{2\sigma_u^2}\right) \tag{46}$$

(v) Integrated Correlation

This formula calculates the sum of the independent correlations regarding time, space, vocabulary, and attributes. For normalization, each element is multiplied by an adjustment weight $\alpha_t, \alpha_\ell, \alpha_v, \alpha_u$.

$$\bar{c}_j = \alpha_t c_j^{(t)} + \alpha_\ell c_j^{(\ell)} + \alpha_v c_j^{(v)} + \alpha_u c_j^{(u)} \tag{47}$$

3.8 Knowledge Mining

The implementation introduces:

$$K = \mathcal{M}(C_B) \tag{48}$$

(i) Rule Extraction

$$v^{(A)} \Rightarrow v^{(B)} \tag{49}$$

(ii) Pattern Extraction

$$d(r_i^{(A)}, r_j^{(A)}) = \|r_i^{(A)} - r_j^{(A)}\| \tag{50}$$

(iii) Prototype Extraction

$$p_k = \arg\max_{j \in P_k} \bar{c}_j \tag{51}$$

3.9 Current-State Retrieval and Prediction

Retrieval and Prediction are implemented as follows:

(i) Retrieval

$$A^* = \sigma_A^*(q^*; \theta) \tag{52}$$

$$r^* = \Gamma_A(A^*) \tag{53}$$

$$s_j = \mathrm{Sim}(r^*, r_j^{(A)}) \tag{54}$$

$$\mathcal{N}_k(r^*) = \arg\mathrm{top}_k \left\{\mathrm{Sim}(r^*, r_j^{(A)})\right\} \tag{55}$$

(ii) Prediction

$$\hat{y} = \frac{\sum_{j \in \mathcal{N}_k} s_j y_j}{\sum_{j \in \mathcal{N}_k} s_j} \tag{56}$$

PoS(ISGC2026)001

4. Experiment

This section describes the experiments conducted to evaluate the effectiveness and feasibility of the proposed method based on the implementation presented in Section 3. The objective of the experiment is to demonstrate that the components defined in the abstract framework can be concretely implemented using real meteorological and landslide data, and that semantic correlation structures between these heterogeneous data sets can be constructed and connected to knowledge representations and retrieval mechanisms.

In this experiment, set A is defined as daily meteorological observation data obtained from the GHCN-Daily dataset provided by NOAA [4], specifically from the Juneau Airport station. The observation period spans from September 1, 1936, to March 12, 2026, and a total of 29,750 daily records are stored in PostgreSQL. Each record contains attributes such as observation date, precipitation, maximum temperature, minimum temperature, average wind speed, and average humidity. Set B is defined as landslide events extracted from the Southeast Alaska landslide dataset provided by the USGS [5]. Events corresponding to the Juneau region are selected based on geographic coordinates and regional attributes. As a result, a total of 29 landslide events are used in the experiment. Each event includes attributes such as occurrence time, location, landslide type, damage description, number of affected buildings, and number of casualties. Therefore, the experiment applies the proposed method to two asymmetric sets: a fine-grained, high-frequency meteorological event set A and a coarse-grained, low-frequency landslide event set B.

The experimental results show that the implementation successfully generates semantic correlation structures for all 29 landslide events. For each $b_j \in B$, the implementation constructs the corresponding subset A_j, the aggregated meteorological representation $r_j^{(A)}$, the landslide representation $r_j^{(B)}$, and the component-wise correlation vector c_j. The results are illustrated in Figure 1. Each landslide event is represented not by a single meteorological observation but by a subset of multiple observations. This confirms that the proposed method captures the relationship between coarse-grained events and fine-grained precursor data as subset-level correspondences. For example, for multiple landslide events that occurred between December 1 and 2, 2020, high values of 24-hour precipitation and 72-hour cumulative precipitation were observed, along with prolonged rainfall duration. In one mudslide case, the 24-hour precipitation reached 125.2 mm, the 72-hour cumulative precipitation reached 143.5 mm, and the rainfall duration lasted for four days, resulting in an overall correlation score of 0.90. Additionally, multiple landslides occurring within the same rainfall event exhibited similarly high correlation scores. Furthermore, for landslide events that occurred on October 27, 2017, in the Juneau region, the 72-hour cumulative precipitation was 87.4 mm, and the average humidity was 96.0%, resulting in an overall correlation score of approximately 0.87.

Additionally, when the correlation is high (e.g., greater than 0.675), the data consists of entries with similar occurrence dates. For example: 2020-12-01 and 2020-12-02 have 6 entries each; 2017-10-27 has 4 entries; and 2019-10-06 has 5 entries. However, when the correlation falls below 0.675, the data consists of entries with different occurrence dates. The three conditions for causality are: correct temporal order, the existence of a correlation, and the absence of a third factor. This method aggregates events b_j that occurred before a specific event a_j and derives a correlation. This method does not explicitly establish a causal relationship. However, when a positive correlation

PoS(ISGC2026)001

exists but is weak (e.g., below 0.675), it can be interpreted as suggesting the possibility of another common cause underlying the relationship. Note that 0.675 is merely a guideline for this situation and is not a threshold.

```
landslide=# SELECT landslide_id, occurred_at, semantic_label, severity_class,
       prcp_24h_mm, prcp_72h_mm, rainy_days_72h, overall_correlation
FROM sem_bridge.correlation_structure
ORDER BY overall_correlation DESC, occurred_at DESC;
 landslide_id |     occurred_at     |     semantic_label     | severity_class |      prcp_24h_mm       |      prcp_72h_mm       | rainy_days_72h | overall_correlation
--------------+---------------------+------------------------+----------------+------------------------+------------------------+----------------+---------------------
           17 | 2020-12-01 12:00:00 | mudslide_type          | localized      |  125.2000000000000000  |  143.5000000000000000  |              4 |                 0.9
           20 | 2020-12-02 12:00:00 | generic_landslide_type | building_damage |   37.1000000000000000  |  178.3000000000000000  |              4 |                0.89
           18 | 2020-12-02 12:00:00 | mudslide_type          | road_impact    |   37.1000000000000000  |  178.3000000000000000  |              4 |                0.89
            7 | 2017-10-27 12:00:00 | mudslide_type          | road_impact    |   51.1000000000000000  |   87.4000000000000000  |              3 |                0.87
           19 | 2020-12-02 12:00:00 | generic_landslide_type | localized      |   37.1000000000000000  |  178.3000000000000000  |              4 |                0.85
           16 | 2020-12-01 12:00:00 | generic_landslide_type | localized      |  125.2000000000000000  |  143.5000000000000000  |              4 |                0.85
           15 | 2020-12-01 12:00:00 | generic_landslide_type | localized      |  125.2000000000000000  |  143.5000000000000000  |              4 |                0.85
            9 | 2017-10-27 12:00:00 | mudslide_type          | localized      |   51.1000000000000000  |   87.4000000000000000  |              3 |                0.83
            6 | 2017-10-27 12:00:00 | generic_landslide_type | localized      |   51.1000000000000000  |   87.4000000000000000  |              3 |                0.78
            8 | 2017-10-27 12:00:00 | generic_landslide_type | localized      |   51.1000000000000000  |   87.4000000000000000  |              3 |                0.78
            4 | 2015-01-21 12:00:00 | generic_landslide_type | localized      |   50.3000000000000000  |   95.7000000000000000  |              4 |                0.78
           27 | 2024-07-14 11:00:00 | generic_landslide_type | building_damage |   31.5000000000000000  |   71.4000000000000000  |              4 |               0.715
           26 | 2023-08-12 12:00:00 | generic_landslide_type | building_damage |   46.2000000000000000  |   52.8000000000000000  |              4 |               0.715
           24 | 2022-09-26 18:00:00 | generic_landslide_type | building_damage |   56.4000000000000000  |   70.6000000000000000  |              4 |               0.715
           12 | 2019-10-06 12:00:00 | generic_landslide_type | road_impact    |   27.9000000000000000  |   44.1000000000000000  |              3 |               0.715
           14 | 2019-10-06 12:00:00 | generic_landslide_type | road_impact    |   27.9000000000000000  |   44.1000000000000000  |              3 |               0.715
           11 | 2019-10-06 12:00:00 | generic_landslide_type | road_impact    |   27.9000000000000000  |   44.1000000000000000  |              3 |               0.715
           10 | 2019-10-06 12:00:00 | generic_landslide_type | road_impact    |   27.9000000000000000  |   44.1000000000000000  |              3 |               0.715
           29 | 2024-07-14 12:00:00 | generic_landslide_type | localized      |   31.5000000000000000  |   71.4000000000000000  |              4 |               0.675
           28 | 2024-07-14 12:00:00 | generic_landslide_type | localized      |   31.5000000000000000  |   71.4000000000000000  |              4 |               0.675
           13 | 2019-10-06 12:00:00 | generic_landslide_type | localized      |   27.9000000000000000  |   44.1000000000000000  |              3 |               0.675
            3 | 2009-11-14 12:00:00 | generic_landslide_type | localized      |   32.0000000000000000  |   62.0000000000000000  |              4 |               0.675
            2 | 2009-11-14 06:30:00 | mudslide_type          | building_damage |   32.0000000000000000  |   62.0000000000000000  |              4 |               0.675
           25 | 2023-01-25 12:00:00 | rockfall_type          | road_impact    |   10.9000000000000000  |   34.2000000000000000  |              4 |                0.57
           23 | 2022-04-10 12:00:00 | generic_landslide_type | road_impact    | 0.0000000000000000000  |  1.8000000000000000000  |              2 |                0.57
           22 | 2022-04-06 12:00:00 | generic_landslide_type | road_impact    | 0.3000000000000000000  |  5.4000000000000000000  |              4 |                0.57
           21 | 2022-01-19 12:00:00 | generic_landslide_type | localized      |   15.2000000000000000  | 15.2000000000000000000  |              2 |                0.57
            5 | 2017-02-15 12:00:00 | generic_landslide_type | building_damage |    8.1000000000000000  |   37.6000000000000000  |              4 |                0.57
            1 | 2002-07-10 12:00:00 | generic_landslide_type | road_impact    |    5.3000000000000000  |  9.6000000000000000000  |              2 |                0.57
(29 rows)
```

Figure 1: The Created Semantic Correlation Structures.

Based on the experimental results, the effectiveness and feasibility of the proposed method are discussed as follows. First, the results confirm that the proposed method can consistently connect two sets with different granularities and element sizes in real-world data. The experiment demonstrates that subsets A_j can be constructed for each landslide event and transformed into comparable feature representations. Second, the results confirm that the proposed method can generate semantically interpretable correlation structures rather than mere numerical similarity. Features such as precipitation, rainfall duration, and moisture conditions are associated with the semantic descriptions of landslide events, enabling the method to describe which precursor states correspond to which types of events. Third, the results confirm that the constructed correlation structures can be utilized for retrieval and knowledge mining. The implementation enables rule extraction and similar-case retrieval based on the correlation structure database C_B, demonstrating that the proposed method transforms correlation results into reusable knowledge representations. Furthermore, the feasibility of the proposed method is confirmed. All processes are implemented using standard PostgreSQL tables and SQL queries without requiring specialized systems. The fact that correlation structures are successfully generated for all landslide events indicates that the proposed method can operate as a practical processing framework on real databases. However, several limitations exist. First, the meteorological data are derived from a single observation station with daily resolution, which limits spatial and temporal granularity. Second, the correlation functions used in this experiment are rule-based and designed for interpretability, rather than optimal predictive performance. Third, the experiment is conducted on a limited regional dataset. Nevertheless, the primary objective of the proposed method is not to maximize predictive accuracy for a specific domain, but to construct semantic correlation structures between heterogeneous data sets and represent them as interpretable knowledge. From this perspective, the experiment demonstrates the fundamental validity and extensibility of the proposed method.

PoS(ISGC2026)001

5. Conclusion

This paper has proposed a Semantic Correlation and Knowledge Mining method that identifies semantic correlation structures between two sets A and B with different granularities and numbers of elements, and performs semantic interpretation and knowledge representation of coarse-grained events. The key feature of the proposed method is the introduction of a framework in which each element $b_j \in B$ is explained not by a single corresponding element, but by a subset $A_j \subseteq A$ constructed from set A. This enables the proposed method to represent relationships between heterogeneous data sets with different granularities not at the level of individual elements, but at the level of subset structures. Furthermore, the proposed method represents each element in a semantic space composed of time, space, and vocabulary, and explicitly constructs semantic correspondences between heterogeneous data sets by computing correlations between aggregated subset representations and coarse-grained event representations. In addition, the proposed method transforms the resulting correlation structures into reusable knowledge representations in the form of rules, patterns, and concepts. In the application to the disaster prevention domain, the proposed method constructs subsets of meteorological observations corresponding to each landslide event and generates semantic correlation structures between these subsets and the events. The results confirm that the proposed method can describe relationships between heterogeneous data sets in an interpretable manner and connect them to retrieval and knowledge mining processes. The proposed method is designed as a domain-independent framework and is not limited to disaster prevention. It can be applied to a wide range of domains that require analysis of relationships between heterogeneous event sets with different granularities, including business, education, and human behavior understanding.

References

[1] B. Wang et al., "Spatiotemporal analysis and threshold modeling of rainfall-induced geological disasters in Anhui Province," *Front. Earth Sci.*, vol. 13, 2025. doi: 10.3389/feart.2025.1541242.

[2] S. Segoni, A. Rosi, D. Lagomarsino, R. Fanti, and N. Casagli, "Brief communication: Using averaged soil moisture estimates to improve the performances of a regional-scale landslide early warning system," *Nat. Hazards Earth Syst. Sci.*, vol. 18, no. 3, pp. 807–812, 2018. doi: 10.5194/nhess-18-807-2018.

[3] Y. Hayashi, Y. Kiyoki, Y. Harada, K. Makino, and S. Kaneoya, "A spatio-temporal and categorical correlation computing method for inductive and deductive data analysis," in *Frontiers in Artificial Intelligence and Applications, Vol. 380: Information Modelling and Knowledge Bases XXXV*, pp. 287–296, IOS Press, 2024. doi: 10.3233/FAIA231162.

[4] M. J. Menne et al., "Global Historical Climatology Network - Daily (GHCN-Daily), Version 3," NOAA National Centers for Environmental Information, 2012. doi:10.7289/V5D21VHZ.

[5] U.S. Geological Survey (USGS), "Landslide Hazards Program – Data and Tools," https://www.usgs.gov/programs/landslide-hazards/data, accessed Mar. 2026.

PoS(ISGC2026)001