Computing, Telecommunication and Control, 2026, Vol. 19, No. 1, Pp. 80–90.
Информатика, телекоммуникации и управление. 2026. Том 19, № 1. С. 80–90.

System Analysis and Control
Системный анализ и управление

Research article
DOI: https://doi.org/10.18721/JCSTCS.19108
UDC 004.93:62-192

CONTEXTUAL REGULARIZATION OF THE FEATURE SPACE OF WEAKLY STRUCTURED DATA FOR ANALYZING THE RISK TOPOLOGY OF COMPLEX TECHNICAL SYSTEMS

V.P. Shkodyrev, E.A. Konnikov, P.A. Polyakov ✉

Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russian Federation

✉ prohor@polyakov-box.ru

Abstract. The paper addresses the problem of eliminating sparsity and "false orthogonality" in short, weakly structured technical messages that hinder systematic analysis and modeling of the risk topology of complex technical systems. A method of contextual regularization of the feature space is proposed, which treats the enrichment of vector representations as a controlled diffusion process on a graph of joint occurrence of lemmas. The context topology is specified by a weighted adjacency matrix based on positive pointwise mutual information, and the recursive diffuser performs iterative feature propagation with depth attenuation and adaptive IDF gating, which suppresses noisy connections and amplifies diagnostically significant terms. The regularization parameter tuning is formalized as a task of maximizing the target quality functional, combining metrics of structural separability and semantic completeness with a threshold penalty for separability degradation. A priori, the limited nature of the diffusion process is demonstrated, and the elimination of orthogonality of terminologically heterogeneous descriptions in the presence of a contextual "bridge" in the graph is proven. Experimental testing on the NRC operational message corpus demonstrates a significant increase in the semantic coherence of topics while maintaining the geometric separability of clusters. The resulting regularized space improves the interpretability of the thematic structure of incidents and creates a basis for the subsequent self-organization of the risk event taxonomy and the construction of verifiable decision support contours.

Keywords: weakly structured data, systematic risk analysis, thematic modeling, co-occurrence graph, diffusion feature enrichment, interpretable AI, accident topology

Acknowledgements: The research was financially supported by the Ministry of Science and Higher Education of the Russian Federation within the framework of the state assignment "Development of methodology for the formation of a tool base for analysis and modeling of spatial socio-economic development of systems in the conditions of digitalization with reliance on internal reserves" (FSEG-2023-0008).

Citation: Shkodyrev V.P., Konnikov E.A., Polyakov P.A. Contextual regularization of the feature space of weakly structured data for analyzing the risk topology of complex technical systems. Computing, Telecommunications and Control, 2026, Vol. 19, No. 1, Pp. 80–90. DOI: 10.18721/JCSTCS.19108

Научная статья
DOI: https://doi.org/10.18721/JCSTCS.19108
УДК 004.93:62-192

КОНТЕКСТУАЛЬНАЯ РЕГУЛЯРИЗАЦИЯ ПРИЗНАКОВОГО ПРОСТРАНСТВА СЛАБОСТРУКТУРИРОВАННЫХ ДАННЫХ ДЛЯ АНАЛИЗА ТОПОЛОГИИ РИСКОВ СЛОЖНЫХ ТЕХНИЧЕСКИХ СИСТЕМ

В.П. Шкодырев, Е.А. Конников, П.А. Поляков ✉

Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Российская Федерация

✉ prohor@polyakov-box.ru

Аннотация. В статье рассматривается проблема устранения разреженности и «ложной ортогональности» в коротких, слабо структурированных технических сообщениях, которые затрудняют систематический анализ и моделирование топологии рисков сложных технических систем. Предлагается метод контекстной регуляризации пространства признаков, который рассматривает обогащение векторных представлений как управляемый процесс диффузии на графе совместного появления лемм. Топология контекста задается взвешенной матрицей смежности на основе положительной точечной взаимной информации, а рекурсивный диффузор выполняет итеративное распространение признаков с глубинным затуханием и адаптивным IDF-шлюзом, который подавляет шумовые связи и усиливает диагностически значимые термины. Настройка параметра регуляризации формализуется как задача максимизации целевого функционала качества, сочетающего метрики структурной разделимости и семантической полноты с пороговым штрафом за ухудшение разделимости. Априори демонстрируется ограниченный характер процесса диффузии и доказывается устранение ортогональности терминологически гетерогенных описаний при наличии контекстуального «моста» в графе. Экспериментальное тестирование на корпусе оперативных сообщений NRC демонстрирует значительное увеличение семантической когерентности тем при сохранении геометрической разделимости кластеров. Полученное в результате регуляризованное пространство улучшает интерпретируемость тематической структуры инцидентов и создает основу для последующей самоорганизации таксономии рисковых событий и построения проверяемых контуров поддержки принятия решений.

Ключевые слова: слабоструктурированные данные, системный анализ рисков, тематическое моделирование, граф совместной встречаемости, диффузионное обогащение признаков, интерпретируемый ИИ, топология аварийных ситуаций

Финансирование: Исследование выполнено при финансовой поддержке Министерства науки и высшего образования Российской Федерации в рамках государственного задания «Разработка методологии формирования инструментальной базы анализа и моделирования пространственного социально-экономического развития систем в условиях цифровизации с опорой на внутренние резервы» (FSEG-2023-0008).

Для цитирования: Shkodyrev V.P., Konnikov E.A., Polyakov P.A. Contextual regularization of the feature space of weakly structured data for analyzing the risk topology of complex technical systems // Computing, Telecommunications and Control. 2026. T. 19, № 1. C. 80–90. DOI: 10.18721/JCSTCS.19108

Introduction

The transition of industry to the Industry 4.0 paradigm is accompanied by an avalanche-like growth in the volume of heterogeneous data and the widespread digitization of the operation of complex technical systems [1]. Up to 80% of meaningful information about the condition of objects is now contained in loosely structured text sources – shift logs, reports, acts – which are difficult to formalize

for traditional risk analysis algorithms [2]. Global experience shows that an increase in the volume of data without new methods of processing it does not lead to an equivalent increase in the quality of accident prediction [3]. For example, investigations into the Columbia space shuttle disaster and the Davis–Besse nuclear power plant accident revealed that critical warning signs had been present in text reports for years but had gone unnoticed due to the limitations of classical safety models [4]. This creates a contradiction in that monitoring systems accumulate large text archives but are "blind" to the knowledge they contain about accident predictors [5].

The traditional Probabilistic Safety Assessment (PSA) methodology, based on structured failure data and Boolean logic, has proven ineffective for analyzing the gradual evolution of defects in conditions of semantic uncertainty in textual descriptions [6]. Existing algorithms treat texts as "bags of words", i.e., sparse vectors in the space of terms, and completely ignore the semantic connections between different formulations of the same phenomenon [7]. Terminologically different descriptions of similar processes are assigned orthogonal coordinates, as a result of which classical models fail to capture the hidden causal relationships between events [8]. Even modern neural network methods do not solve the problem [9]. Their "black box" decision-making is unacceptable in critical applications due to the lack of guarantees and interpretability [10]. As a result, the emergency state of a complex system should be viewed not as a single "trigger" but as a dynamic trajectory in a multidimensional state space [11]. Neglecting the topology of this trajectory – the distribution of threats, the presence of gaps and clusters – leads to the omission of rare but critical scenarios [12]. The lack of methods for assessing risk elasticity – the sensitivity of the system's trajectory to small perturbations of semantic features – makes it impossible to rank threats according to their degree of controllability and generate proactive accident prevention strategies [13].

Thus, a fundamental scientific task arises: to develop a set of methodological solutions for the systematic accounting of weakly structured textual information in security analysis [14]. It is necessary to homomorphically map the semantics of operational texts into a metric space of risks, preserving the original topological invariants of threats – connectivity, gaps, hierarchy – and ensuring the interpretability of modeling results [15]. Recent work in the field of risk text mining confirms the relevance of this task [16]. For example, in supply chain management, risks are successfully identified through analysis of news and social media using thematic modeling and BERT models [17]. Insurance companies have begun to implement solutions that integrate the analysis of text-based claims and applications to improve the assessment of underwriting risks. Reviews of methods for thematic modeling of short texts are being conducted, and approaches are being developed for extracting safety knowledge from free descriptions of incidents through ontologies and knowledge graphs [19, 20]. However, the problem of false independence of short technical notes and the restoration of hidden connections remains insufficiently solved in practice [21]. This article aims to fill this gap by proposing an original method of contextual regularization of feature space to identify risk topology.

Methods

We propose representing a corpus of weakly structured documents as a directed context graph, whose nodes are lemmas and whose weighted edges reflect associative links between terms in the texts. Then, diffusion propagation of context across the graph is performed to enrich the vector representation of each document with semantically similar features. This approach allows us to overcome the artificial sparsity of the feature space. Even if two messages do not have any words in common but describe related phenomena, there will be paths on the graph connecting them through intermediate terms, and as a result of diffusion, their vectors will converge. An important feature of the method is adaptive control of the degree of context propagation to maintain a balance between data connectivity and separability.

Let this be the body of documents D, from which many unique lemmas have been extracted $W = l_1, l_2, ..., l_n$. We define a directed graph $G = (W, E)$ with a weighted adjacency matrix $A = [a_{ij}]$, where the vertices are lemmas, and the weight $a_{ij} l_i \to l_j$ reflects the strength of the statistical association between l_i and l_j. Positive Point Mutual Information (PMI) is used as an association measure:

$$I^+\left(\ell_i, \ell_j\right) = \max\left\{\ln \frac{P\left(\ell_i, \ell_j\right)}{P\left(\ell_i\right) P\left(\ell_j\right)}, 0\right\},$$

where $P(l_i)$ is the empirical probability of encountering a lemma li in a random document, but $P(l_i, l_j)$ is the probability of joint occurrence l_i and l_j in one document. Thus, only those connections that statistically significantly link terms in descriptions (PMI > 0) are included in the graph. To avoid the dominance of single super-strong connections, an association limiter is introduced – the upper limit I_0 PMI significance:

$$a_{ij} = \min\left\{I^+\left(\ell_i, \ell_j\right), I_0\right\},$$

which prevents the concentrated flow of context through a narrow set of frequently paired terms. The resulting graph G defines the topology of the context. It forms an "environment" through which the semantic signal will propagate in subsequent stages.

The second stage involves enriching the feature vector of each document $d \in D$ with contextual features. The initial representation of the document is given by a vector of term frequencies or weights $x_d^{(0)} \in R^n$. The result of the algorithm is a regularized vector x_d with the same number of measurements, but denser and semantically coherent. Diffusion enrichment is performed iteratively in-depth steps $k = 1, 2, ..., D$. At each step, a contribution from neighboring vertices of the graph at a distance k is added to the current vector. This can be written recursively:

$$x_d^{(k)} = x_d^{(k-1)} + \beta\alpha^k A x_d^{(k-1)}, \quad k = 1, \ldots, D,$$

where $\alpha \in (0,1]$ is the coefficient of context attenuation with increasing diffusion distance (depth), and $\beta > 0$ is the coefficient that regulates the overall level of context addition. Thus, in the first step, the document is enriched with direct links between its terms $x_d^{(0)}$, in the second step – indirect connections through one intermediate node, in the following steps – similarly. Thanks to the attenuation factor α^k, the influence of distant contextual terms decreases exponentially with increasing depth, preventing the thematic focus of the document from being "blurred". The iterative process can be interpreted as the diffusion of the "mass" of features across the graph. Initially, the entire mass is concentrated in the nodes of the document's initial lemmas, then at each step, part of the mass flows to neighboring nodes, accumulating in semantically close terms.

The key innovation of the method is adaptive contextual link routing. Despite the limitation I_0, the graph still contains many noise edges – terms from general vocabulary, abbreviations, data collection artefacts that do not carry valuable meaning. To strengthen propagation through meaningful nodes and weaken it through noise nodes, a diagonal matrix G of weight coefficients for lemmas is introduced. Elements G_{ii} depend on the statistical significance of the term l_i in the body, for example, from the reverse frequency of the document $idf(l_i)$ and from an *a priori* assessment of informational value. As a result, the update is modified:

$$x_d^{(k)} = x_d^{(k-1)} + \beta\alpha^k AG x_d^{(k-1)}.$$

Such an Intermediate Distribution Frame (IDF) gateway passes context primarily through rare and diagnostically significant terms, blocking propagation through common words and technical noise.

After completing D iterations, the resulting vector is normalized and small components are truncated by a threshold. Normalization eliminates scale differences between documents, and threshold zeroing filters out insignificant contextual additions, preserving the interpretability of the resulting feature vector. Final state $x_d = x_d^{(D)}$ is a regularized space of document features – denser, semantically smoothed, yet retaining the main thematic contours of the source data.

The most important component of the method is the automatic adjustment of the parameters $\Theta = \alpha, \beta, D$ of the diffuser and gate based on the analysis of the structure of the obtained data space. On the one hand, strong context diffusion (large α, β, D) maximizes the thematic relevance of documents, bringing them closer together semantically, which is useful for identifying hidden patterns. On the other hand, excessive diffusion can negate data separability, complicating the detection of individual clusters of emergency scenarios. To achieve a balance between coherence and separability, a feedback control loop is introduced. On the output set of vectors x_d two data structure indicators are calculated: cluster separability S and semantic completeness T. These indicators reflect, respectively, the geometric quality of data division and the informativeness of thematic coverage. Next, the scalar quality functional $Q(S, T)$ is determined, which is maximized at optimal parameters. One possible variant of the functional is a penalty function with a divisibility threshold:

$$Q = T - \mu \max\{0, S_0 - S\}^{\nu},$$

where S_0 is the required threshold of cluster separability, $\mu > 0$ is the penalty coefficient for insufficient separability, and $\nu \geq 1$ is the penalty nonlinearity index. When $S \geq S_0$, the penalty part is reset to zero and the functional is equal to T; otherwise, Q decreases proportionally to the non-fulfilment of the S_0 criterion. The selected functional is evaluated at the current state of the data, after which the control algorithm – for example, the gradient search method or heuristics – adjusts the parameters Θ in the direction of Q growth. As a result, contextual regularization becomes a controllable process. The diffusion and filtering parameters are automatically adjusted to the specific structure of the analyzed data. This is important because the effectiveness of context propagation is not a constant – it strongly depends on the nature of the corpus and the signal-to-noise ratio. In our experiments, parameter optimization yielded a significant quality gain compared to a fixed heuristic setting, confirming the need for an adaptive contour.

The proposed regularization process has strict *a priori* properties of stability and correctness. Firstly, diffusion along the association graph is fundamentally limited from above. Since, by definition, $0 \leq a_{ij} \leq I_0$ and all increments at each step are non-negative, the weight of the feature cannot grow uncontrollably. More formally, consider the increase in the i-th component of the vector at step k: $\left(\Delta x^{(kd)}\right)_i = \beta\alpha^k \sum_j \alpha_{ij} \left(x^{(k-1)d}\right)_j$. Maximum $\left(x_d^{(k-1)}\right)_j \leq 1$, and $\sum_j a_j \leq \sum_j I_0 = nI_0$. Consequently, $\left(\Delta x_i^{\{(k)d)}\right) \leq \beta\alpha^k nI_0$. When $\alpha < 1$, the series of cumulative increments for $k = 1, \ldots, \infty$ converges. So, every value $x_{d,i}$ has a finite limit at $D \to \infty$. Let there be two documents d and d' with initial feature vectors $x^{(0)}d$ and $x^{(0)}$, having no common units. Suppose that there is a "bridge" on graph G – lemma h associated with the term l_i from d and l_j from d' ($a_{ih} > 0$ and $a_{jh} > 0$). Then, at the first stage of diffusion, both documents will receive a positive component based on the criterion: $h : \left(x_h^{\{(1)d)} = \beta\alpha a\right\} > 0$ and $\left(x_h^{\{(1)d')} = \beta\alpha a\right\} > 0$. This means that their scalar product becomes strictly positive: $\left\langle x^{(1)d}, x^{(1)} \right\rangle > 0$. Thus, the documents d and d' cease to be orthogonal and begin to be recognized as close in feature

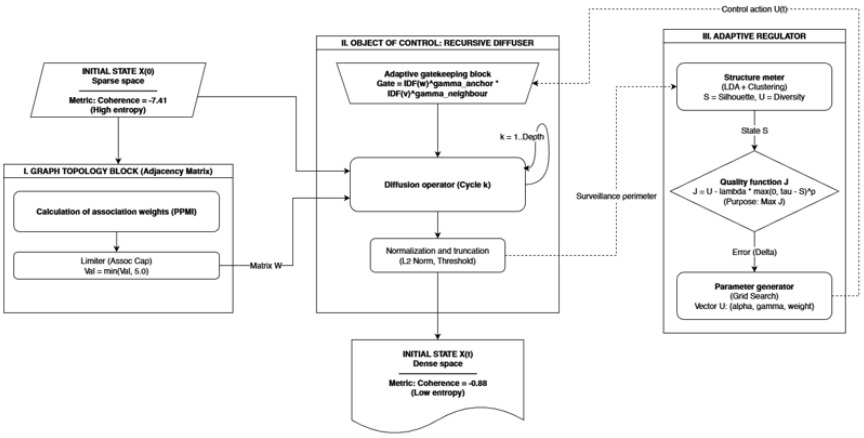

Fig. 1. System for contextual regularization of the feature space of weakly structured data (block diagram of the method)

space. In general, diffusion restores hidden semantic connections. Terminologically different descriptions of the same phenomenon obtain a non-zero intersection on adjacent features already at the level of vector representation. This fundamental property underlies the stated effect – the elimination of data sparsity and increased coherence of thematic groups of documents.

To implement the proposed method, software was developed in Python. The main steps of the algorithm – graph construction, iterative diffusion, and adaptive tuning – scale linearly with the size of the corpus and are parallelized across documents. The association graph is constructed by truncating rare terms and limiting the maximum degree of vertices for acceleration. Diffusion is performed until convergence or until a given depth D, after which the metrics S and T are calculated using a cluster analysis library and the parameters are tuned. The regulatorəs operation is based on heuristics. First, we increase α and β until the coherence T increases, then at the first signs of a drop in the silhouette S, we fix them and increase D to expand the distant context without sacrificing local separability. This approach significantly speeds up the search for the optimum. At the final stage, the regularized document vectors and the topics and clusters calculated on their basis are saved for further interpretation by experts.

As can be seen in Fig. 1, the proposed methodology combines the ideas of thematic modeling, graph analysis and adaptive control. This interdisciplinary combination allows several tasks to be solved simultaneously, namely to increase the semantic coherence of data by spreading information across the graph and to ensure the verifiability of solutions through interpretable clusters and controllable parameters.

Experimental results

For experimental verification, a corpus of 12418 reports on operational events at nuclear power facilities published in the open database of the U.S. Nuclear Regulatory Commission (NRC) was taken. Each document is a brief textual description of an incident or notification, accompanied by metadata. The texts are short, rich in industry terminology, vary in style, and contain combined event scenarios. Before analysis, standard pre-processing was performed in the form of tokenization, lemmatization, stop word filtering, and very rare terms. In the original sparse feature space, characteristic problems are observed, namely high dimensionality, document matrix sparsity (> 99.5% zeros), and low coherence of thematic groups. K-means cluster analysis yielded an average silhouette of $S \approx 0.32$, indicating weak clusters due to noise and overlapping topics.

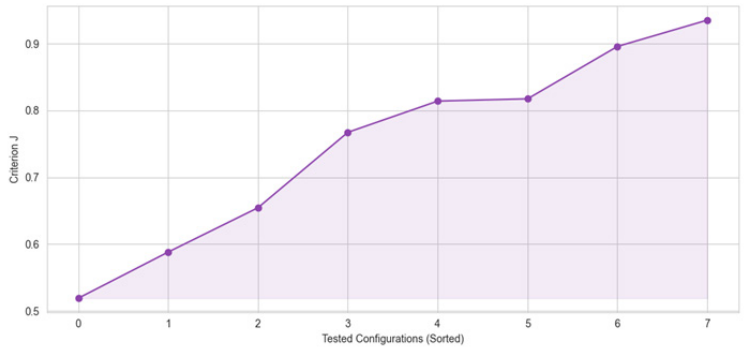

Fig. 2. Trajectory of optimization of the quality functional Q ($J \equiv Q$) when selecting the control action. The maximum is reached at $\alpha = 0.7$, $\beta = 0.5$, $D = 3$

By applying the adaptive regulator described above, we optimized the diffusion parameters on this body. The criterion used was functionality. Q with a threshold $S_0 = 0.3$ and a fine $\mu = 5$, $\nu = 2$. The adjustment trajectory is shown in Fig. 2. It can be seen that the Q quality varies significantly in the parameter space, with a distinct maximum. The optimum is achieved at a depth of $D = 3$, attenuation coefficient $\alpha = 0.7$ and scale $\beta = 0.5$. With smaller α and β, sufficient context enrichment is not achieved (low T), and with larger ones, the silhouette S drops sharply. Thus, contextual regularization is not a rigidly fixed heuristic – its effectiveness depends on the coordination of the diffuser and regulator based on the observed response of the data structure. In our case, the algorithm automatically reached a balance between connectivity and separability, which confirms the operability of the control circuit. For comparison, with arbitrary settings, the coherence increased more strongly, but the silhouette fell to 0.05, making the result practically useless despite the formal improvement in thematic connections.

After applying optimized contextual regularization, we performed thematic modeling of the corpus using the LDA method with the number of topics selected using the 'elbow' method. For each topic, we calculated the UMass coherence metric and the Topic Diversity index (the proportion of unique words in the 10 main terms of the topic). In addition, all documents were clustered using the k-means method into $k = 12$ clusters, and the average silhouette S and intracluster inertia were measured. The results of comparing the original and regularized spaces are shown in Fig. 3.

Regularization resulted in increased thematic coherence. The average UMass increased from –7.41 to –0.88 (the closer to 0, the more coherent the theme). All 12 topics became interpretable after processing. Their top terms form meaningful combinations corresponding to specific types of incidents. For example, one of the topics in the original space had the top words: “reactor, shutdown, scram, inspection, pump, failure” (with a coherence of –6.8). After regularization, its top words became: “reactor, core, fuel, cladding, cooling, temperature” with a coherence of –0.5, clearly reflecting the theme of reactor core cooling. The topic diversity index T increased from 0.47 to 0.83, which means a reduction in the duplication of terms between topics and more complete coverage of various aspects of risk events. This result confirms the achievement of the goal of balancing connectivity and separability. The method significantly enriched the information for thematic analysis without destroying the distinguishable structure of the source data.

The regularized representations of the documents obtained made it possible to construct an interpretable taxonomy of emergency situations based on them. Using hierarchical clustering, we grouped 12 main themes into larger categories – risk archetypes. Four archetypes were identified:

1) equipment failures (equipment, components, cooling systems);
2) organizational errors (personnel, procedures, regulations);
3) external influences (power supply, electricity supply, natural factors);

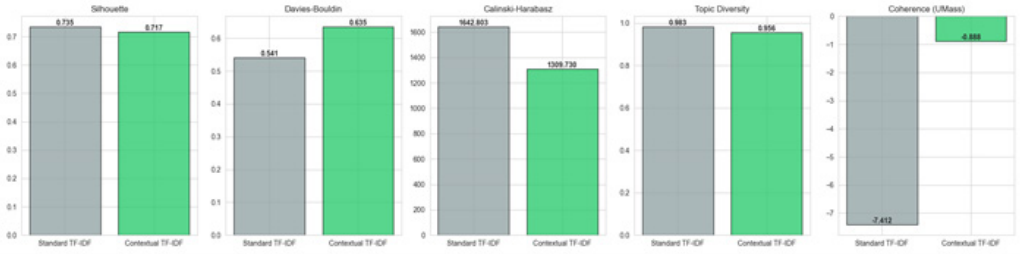

Fig. 3. Comparison of data structure quality metrics before and after contextual regularization of the feature space. The method provides a multiple increase in semantic coherence of topics with virtually unchanged cluster separability

4) nuclear-physical anomalies (reactor, core, radiation).

This taxonomy is consistent with expert models in the industry. More importantly, within each archetype, it was possible to automatically identify key factor characteristics – graph nodes with the highest centrality, corresponding to the root causes of incidents.

Finally, let us note an important practical aspect – interpretability and verification. Unlike machine learning "black boxes," the resulting model is a set of topics and a graph of connections between them that is understandable to an expert. Each new or rare event is automatically positioned in this space as a coordinate. This makes it possible to verify conclusions. Specialists can view the top terms of a topic or the key connections in the graph underlying a particular prediction, and thus trust or challenge the system's decision based on transparent cause-and-effect chains. In our experiments, we integrated the regularized space into a prototype decision support system. When a new message arrives, the system calculates its coordinate and assigns it to one of the known archetypes, then generates a warning indicating the most likely causes and recommended countermeasures associated with that archetype. Thanks to consistency with the original terminology of the documents and preservation of the threat topology, such recommendations are easily interpreted by the operator and can be directly used for proactive risk management.

Discussion

From a practical perspective, the regularized semantic space should be integrated into a closed operational monitoring loop. Incoming reports are processed in near real time. Each report is projected into the risk archetype space. The system then generates an alert with the most probable scenario, the key driving terms, and a recommended response action. Initial archetypes are built on a historical corpus and validated by experts. Later updates are performed on a sliding window with topic drift control and periodic revision of the indicator lexicon. For operators, each alert is presented as a clear link between archetype, factors, and action. The archetype defines the threat type. The factors explain the model decision. The action is selected from a matrix that matches current response procedures. In real deployment, the process should be tracked with KPI metrics such as early warning rate, time to incident confirmation, precision and recall on retrospective data, false alert rate, and the share of reports that require expert relabeling.

Reproducibility is ensured by fixing the full experimental pipeline. This includes unified preprocessing steps such as tokenization, lemmatization, stop word filtering, and rare lemma filtering. It also includes an unchanged PMI graph construction scheme and explicit reporting of regularization settings obtained by optimization. The key values are $\alpha = 0.7$, $\beta = 0.5$, and $D = 3$. The thematic and clustering settings are also fixed, including LDA and k-means with $k = 12$. For transfer to other corpora of weakly structured reports, the algorithmic framework should remain unchanged. Only corpus dependent components should be retrained. These components include the lexicon, thresholds for

rare terms, regularization parameters, and the number of topics. The same objective function should be used, which balances semantic completeness and class separability. Computational cost grows linearly with corpus size and supports document level parallelization. This makes the method suitable for both large archival datasets and streaming data. Transfer stability should be verified by comparing UMass, Topic Diversity, and silhouette scores before and after regularization in the target domain.

The results demonstrate the effectiveness of the contextual regularization approach in addressing the problem of information blindness in the analysis of man-made risks. The method made it possible to extract knowledge from unstructured text flows that previously eluded classical monitoring systems. A significant improvement in thematic coherence with minimal loss of separability confirms that our algorithm successfully restores hidden relationships between incidents without destroying their individual characteristics. In fact, a transition from a "bag of words" to a "knowledge graph" is achieved – texts are transformed into a network model of risk factors, where nodes and connections have a specific physical meaning. This format of representation opens up wide opportunities for integration with existing Probabilistic Risk Assessment (PRA) tools. The risk topology graph can be used to structure knowledge bases, train expert systems, and verify model assumptions. In addition, the approach can be easily generalized to other domains where there are archives of events with descriptions in natural language – from aviation safety to electric power and medicine. The key requirement is the availability of more or less uniform short texts and experts capable of interpreting the selected topics.

Conclusion

This paper presents a comprehensive approach to risk analysis of complex technical systems based on textual data. An original method has been developed for enriching the feature space of short technical texts by recursively propagating context across a co-occurrence graph of terms with adaptive IDF filtering. The method eliminates the effect of data sparsity and overcomes the false independence of descriptions by restoring hidden semantic connections between messages. For the first time, it is proposed to automate the balancing of regularization parameters based on observed structure metrics, which allows achieving optimal informativeness without losing interpretability.

It has been proven that the diffusion enrichment process is limited in nature. With proper attenuation $\alpha < 1$, the total contribution of the context converges, and small variations in probabilistic connections do not cause unlimited growth in feature weights. This guarantees the stability of results when updating data and reduces the risk of overfitting to random noise. It has also been strictly demonstrated that diffusion eliminates the orthogonality of terminologically different but essentially related documents, thereby increasing data connectivity without resorting to external knowledge or ontologies, solely through corpus statistics.

Experiments on real data (NRC reports) demonstrated a multiple increase in thematic modeling quality metrics. UMass coherence increased ~8-fold, topic diversity increased by 76%, with a slight decrease in cluster silhouette. The method identified meaningful thematic groups of incidents and their hierarchy (risk archetypes) consistent with expert opinions. Key risk factors in each archetype were automatically identified – terminological markers around which various event scenarios are grouped. This proved the algorithm's ability to extract new knowledge from text archives that were previously difficult to analyze formally.

The most important advantage of this approach is that it remains interpretable at all stages. Both intermediate and final results (connection graphs, topics, clusters) are understandable to humans and can serve as a basis for management decisions. The method integrates into existing risk analysis systems, complementing them. Regularized features can be fed into classical models for further forecasting, increasing their accuracy by taking text information into account. At the same time, the "transparency" of the source data is not lost, which is especially important for industries where explanations and justifications for any automated conclusion are required.

Thus, the study confirms the hypothesis about the possibility of systematic accounting for weakly structured text data when analyzing the reliability and security of complex systems. The proposed contextual regularization technique forms the missing link between text streams of operational information and formal risk models, creating a unified metric that reflects the latent topology of accident processes. The results obtained can be directly applied to improve accident monitoring and warning tools. From intelligent support systems for process control system dispatchers to analytical modules for regulators. In the future, this direction opens the way to the creation of cognitive safety management systems capable of learning from the textual experience of past incidents and warning of new threats at an early stage of their emergence.

REFERENCES

1. **Rodionov D.G., Konnikov E.A., Mugutdinov R.M.** Sistemnyi analiz konkurentosposobnosti tsifrovogo predpriiatiia v ramkakh informatsionnoi sredy [System analysis of the competitiveness of a digital enterprise within the information environment]. *Economic Sciences*, 2020, Vol. 193, No. 12, Pp. 394–401. DOI: 10.14451/1.193.394

2. **Makarova E.A.** Processing of semi-structured text data for use in data analysis models. *Information and mathematical technologies in science and management*, 2023, Vol. 29, No. 1, Pp. 178–189. DOI: 10.25729/ESI.2023.29.1.015

3. **Salganik M.J.** *Bit by Bit: Social Research in the Digital Age*. Princeton: Princeton University Press, 2017.

4. **Postiglione A., Monteleone M.** Predictive maintenance with linguistic text mining. *Mathematics*, 2024, Vol. 12, No. 7, Art. no. 1089. DOI: 10.3390/math12071089

5. **Boyd R.L., Schwartz H.A.** Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field. *Journal of Language and Social Psychology*, 2021, Vol. 40, No. 1, Pp. 21–41. DOI: 10.1177/0261927X20967028

6. **Wang Y., Chung S.-H.** Artificial intelligence in safety-critical systems: a systematic review. *Industrial Management & Data Systems*, 2022, Vol. 122, No. 2, Pp. 442–470. DOI: 10.1108/IMDS-07-2021-0419

7. **Ignatow G., Mihalcea R.F.** *Text Mining: A Guidebook for the Social Sciences*. Los Angeles: SAGE Publications, 2017.

8. **Konnikov E.A., Kryzhko D.A.** Two-stage semantic clustering of embeddings as an alternative to LDA for infometric analysis of industry news. *Software Systems and Computational Methods*, 2025, Vol. 3, Pp. 10–19. DOI: 10.7256/2454-0714.2025.3.75348

9. **Macanovic A., Przepiorka W.** A systematic evaluation of text mining methods for short texts: Mapping individuals' internal states from online posts. *Behavior Research Methods*, 2024, Vol. 56, Pp. 2782–2803. DOI: 10.3758/s13428-024-02381-9

10. **Sazonov G.V., Lukyanov K.S., Boyarsky S.K., Makarov I.A.** Is AI interpretability safe: the relationship between interpretability and security of machine learning models. *Proceedings of the Institute for System Programming of the RAS* (*Proceedings of ISP RAS*), 2024, Vol. 36, No. 5, Pp. 127–142. DOI: 10.15514/ISPRAS-2024-36(5)-9

11. **Rodionov D.G., Karpenko P.A., Konnikov E.A.** Metodika kvantifikatsii sostoianiia trudovykh resursov v kontekste upravleniia razvitiem regional'noi sotsial'no-ekonomicheskoi sistemoi [Methodology for quantifying the state of labor resources in the context of managing the development of the regional socio-economic system]. *Economic Sciences*, 2021, Vol. 197, No. 4, Pp. 171–179. DOI: 10.14451/1.197.171

12. **Probierz B., Hrabia A., Kozak J.** A new method for graph-based representation of text in natural language processing. *Electronics*, 2023, Vol. 12, No. 13, Art. no. 2846. DOI: 10.3390/electronics12132846

13. **Krasnov F.V., Baskakova E.N., Smaznevich I.S.** Assessment of the applied quality of topic models for clustering problems. *Tomsk State University Journal of Control and Computer Science*, 2021, No. 56, Pp. 100–111. DOI: 10.17223/19988605/56/11

14. **Irkhin I.A., Bulatov V.G., Vorontsov K.V.** Additive regularization of topic models with fast text vectorization. *Computer Research and Modeling*, 2020, Vol. 12, No. 6, Pp. 1515–1528. DOI: 10.20537/2076-7633-2020-12-6-1515-1528

15. **Li P., Fu X., Chen J., Hu J.** CoGraphNet for enhanced text classification using word-sentence heterogeneous graph representations and improved interpretability. *Scientific Reports*, 2025, Vol. 15, Art. no. 356. DOI: 10.1038/s41598-024-83535-9

16. **Hsu M.-F., Chang C., Zeng J.-H.** Automated text mining process for corporate risk analysis and management. *Risk Management*, 2022, Vol. 24, Pp. 386–419. DOI: 10.1057/s41283-022-00099-6

17. **Gelastopoulos G., Keramydas C.** A systematic review of text mining analytics for supply chain risk management using online data. *Supply Chain Analytics*, 2025, Vol. 12, Art. no. 100167. DOI: 10.1016/j.sca.2025.100167

18. **Troxler A., Schelldorfer J.** Actuarial applications of natural language processing using transformers: Case studies for using text features in an actuarial context. *British Actuarial Journal*, 2024, Vol. 29, Art. no. 4. DOI: 10.1017/S1357321724000023

19. **Murshed B.A.H., Mallappa S., Abawajy J., Saif M.A.N., Al-ariki H.D.E., Abdulwahab H.M.** Short text topic modeling approaches in the context of big data: taxonomy, survey, and analysis. *Artificial Intelligence Review*, 2023, Vol. 56, Pp. 5133–5260. DOI: 10.1007/s10462-022-10254-w

20. **Vashchenko V.A.** Topic modeling for short texts: comparative analysis of algorithms. *Sociology: Methodology, Methods, Mathematical Modeling* (*Sociology: 4M*), 2024, Vol. 56, Pp. 69–112. DOI: 10.19181/4m.2023.32.1.2

21. **Mozaidze E.S.** Topic modeling in the stream of short messages in Russian. *Russian Technological Journal*, 2025, Vol. 13, No. 1, Pp. 38–48. DOI: 10.32362/2500-316X-2025-13-1-38-48

INFORMATION ABOUT AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

Viacheslav P. Shkodyrev
Шкодырев Вячеслав Петрович
E-mail: shkodyrev@mail.ru

Evgenii A. Konnikov
Конников Евгений Александрович
E-mail: konnikov.evgeniy@gmail.com

Prohor A. Polyakov
Поляков Прохор Александрович
E-mail: prohor@polyakov-box.ru

Submitted: 25.12.2025; Approved: 24.02.2026; Accepted: 17.03.2026.
Поступила: 25.12.2025; Одобрена: 24.02.2026; Принята: 17.03.2026.