


Article

Intelligent Algorithms for Event Processing and Decision Making on Information Protection Strategies against Cyberattacks

Grigorii Asyaev¹, Alexander Sokolov¹ and Alexey Ruchay^{1,2,*} ¹ Department of Information Security, South Ural State University, Chelyabinsk 454080, Russia² Department of Mathematics, Chelyabinsk State University, Chelyabinsk 454001, Russia

* Correspondence: ran@csu.ru

Abstract: This paper considers the main approaches to building algorithms for the decision support systems of information protection strategies against cyberattacks in the networks of automated process control systems (the so-called recommender systems). The advantages and disadvantages of each of the considered algorithms are revealed, and their applicability to the processing of the information security events of the UNSW-NB 15 dataset is analyzed. The dataset used contains raw network packets collected using the IXIA PerfectStorm software in the CyberRange laboratory of the Australian Cyber Security Centre (Canberra) in order to create a hybrid of the simulation of the real actions and the synthetic behavior of the network traffic generated during attacks. The possibility of applying four semantic proximity algorithms to partition process the data into clusters based on attack type in a distribution control system (DCS) is analyzed. The percentage of homogeneous records belonging to a particular type of attack is used as the metric that determines the optimal method of cluster partitioning. This metric was chosen under the assumption that cyberattacks located “closer” to each other in the multidimensional space have similar defense strategies. A hypothesis is formulated about the possibility of transferring knowledge about attacks from the vector feature space into a semantic form using semantic proximity methods. The percentage of homogeneous entries was maximal when the cosine proximity measure was used, which confirmed the hypothesis about the possibility of applying the corresponding algorithm in the recommender system.



Citation: Asyaev, G.; Sokolov, A.; Ruchay, A. Intelligent Algorithms for Event Processing and Decision Making on Information Protection Strategies against Cyberattacks. *Mathematics* **2023**, *11*, 3939. <https://doi.org/10.3390/math11183939>

Academic Editor: Cheng-Chi Lee

Received: 27 July 2023

Revised: 14 September 2023

Accepted: 15 September 2023

Published: 16 September 2023



Copyright: © 2023 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/>).

Keywords: attack vector; cyberattack; decision support system; automated process control system; predictive information protection; collaborative filtering

MSC: 68M25

1. Introduction

The number of successful company attacks that have caused both financial and other negative consequences has been on the rise throughout 2023. The number of attacks on companies has increased by 27% compared to the previous year, with an average financial loss of USD 1.1 million per attack. The Cybercrime Report, updated on 2 September 2023, states that data breaches cost businesses an average of USD 4.35 million in 2022. Around 236.1 million ransomware attacks occurred globally in the first half of 2022 [1]. AI-powered security algorithms are helping to reduce the time it takes to detect and respond to cyberattacks. Intelligence algorithms are helping to detect cyberattacks in several ways: anomaly detection, behavioral analysis, predictive analytics, incident response, threat intelligence, network traffic analysis, endpoint protection, identity and access management, fraud detection, compliance, and governance [2].

There are a large number of studies devoted to the detection of data anomalies and the classification of the attacks that cause them [3]. The next stage in the development

of intrusion detection systems (IDSs) is the development of predictive information protection systems. Issues related to decision-making algorithms in such systems are poorly researched as evidenced by the limited number of articles on this topic [4]. Therefore, the task of modeling decision support systems which would allow an information security administrator to correctly and promptly choose the optimal information protection strategy to prevent possible incidents or to reduce the negative consequences when they occur (the so-called advisory systems) is very important.

Recommender systems are algorithms that analyze user behavior data to generate personalized recommendations for users. They can be used in various domains, including information security, to provide personalized security advice or to recommend security-related products or services to users.

Currently, there is a sufficient number of scientific works devoted to improving the quality of anomaly detection and identification and the multiclass classification of attacks, which are both within the ACS and the Internet of Things. This is confirmed by works [5–14]. The authors are improving and inventing new approaches and detection standards aimed at improving accuracy, but the issue of further actions in the case of the timely detection of an anomaly is poorly studied. One of the subtasks is to build a recommender system with different information security strategies which, depending on the anomaly/attack found, should recommend to the security officer a set of actions aimed at reducing the degree of damage from a cyberattack. For this reason, this paper investigates the practical approach of building an information security recommendation system.

The possible advantages of using recommender systems in information security include the following:

- Increased user awareness and engagement: by providing personalized security recommendations, users may become more aware of potential security threats and take action to protect themselves.
- Improved decision making: Recommender systems can analyze large amounts of data and identify patterns that may be difficult for humans to detect. This can lead to better decision making and improved security outcomes.
- Enhanced user experience: personalized recommendations can improve the user experience by making security advice more relevant and actionable for individual users.

However, there are also some potential disadvantages to using recommender systems in information security, including the following [15]:

- Privacy concerns: Recommender systems rely on user data to generate recommendations. As such, there may be concerns about the privacy and security of user data.
- Biases in recommendations: the recommendations generated by a recommender system may be biased if the system is trained on biased data or if it relies on outdated or incomplete information.
- Lack of human expertise: while recommender systems can analyze large amounts of data, they may lack the human expertise needed to identify complex security threats or to provide nuanced security advice.

Our goal is to identify and formulate approaches for building recommender systems, to analyze the existing recommender system methods in other fields, and to evaluate their applicability to information security.

The main contribution of this article is the initial minimum number of records for a basic information security decision support system run, characterized by fast speed and variable multilevel partitioning for system construction, which allows us to define the protection strategies for a specific anomaly in the data and for different types of attacks.

The remainder of this article is organized as follows: Section 2 gives a description of previous works on the above topics. Section 3 analyses the construction of the recommender systems used in retail, telecoms, etc., and assesses their applicability in the field of information security. Section 4 describes the data for the experiment, the hypothesis, the metrics, and the construction of the recommender system and how it can be applied in

a production environment. Section 5 discusses the proposed recommender system, and Section 6 concludes this study.

2. Related Works

The research gap in the field of cybersecurity is the lack of effective and efficient methods for detecting and responding to cyberattacks, particularly in the context of the Internet of Things (IoT). Existing approaches often rely on technical and operational measures, such as firewalls and intrusion detection systems, which can be insufficient in detecting and mitigating modern cyberthreats.

To address this gap, the authors of [5] propose an improved attack identification process that aggregates technical and organizational security metrics and detection sources. This approach can help identify cyberattacks at an earlier stage and can provide a more comprehensive understanding of the scale of an attack. However, the authors do not address the limitations of their approach, such as the potential for false positives or the need for further action after an anomaly is detected.

Another approach to addressing the research gap is the use of recommender systems in cybersecurity as proposed by the authors of [6]. These systems can provide personalized recommendations for cybersecurity responses and mitigation strategies based on the specific needs and circumstances of an organization. However, the authors note that there is a lack of cybersecurity recommender applications, highlighting the need for further research in this area.

The authors of [7,8] also address the research gap by using the UNSW-NB15 dataset to develop and evaluate intrusion detection frameworks. These frameworks aim to combine a knowledge transfer and resistance to zero-day attacks, including for devices connected to the IoT. However, the authors do not fully address the limitations of their approaches, such as the need for further action after an anomaly is detected or the potential for false positives.

To shed light on the research gap and the limitations of the existing approaches, a table can be used to highlight the methods and shortcomings of other researchers and their proposed methodologies.

The authors of [9] discuss the growing use of smart grids, which incorporate advanced technologies to improve power distribution, but also present their vulnerabilities, particularly in cybersecurity. The study identifies the three layers of the innovative grid network that are vulnerable to cyberattacks: users, the network of smart devices and sensors, and network administrators. To address these vulnerabilities, the authors propose security solutions using various methods, including intrusion detection systems (IDSs) based on deep learning.

However, the study also highlights the limitations and drawbacks of these methods. For example, the authors note that traditional IDSs are not effective against newly emerging threats and that deep-learning-based approaches can be computationally expensive and require large amounts of training data. Moreover, the study acknowledges that there is a lack of standardization in the field of smart grid cybersecurity, which makes it difficult to compare and evaluate different approaches.

In contrast, [10] focuses on the importance of network traffic analysis in ensuring the security of online food-security- and sustainability-related industries. The study proposes an IDS for SCADA networks based on deep learning which can defend against both conventional and SCADA-specific network-based attacks. The proposed approach achieved a high detection accuracy, with the KNN and RF algorithms achieving a near-perfect score of 99.99% and the CNN-GRU model achieving an accuracy of 99.98%.

The Table 1 highlights the strengths and limitations of the methods proposed in [9,10]:

Table 1. The strengths and limitations of the methods proposed in [9,10].

Method	Strengths	Limitations
Traditional IDSs	Effective against known attacks	Not effective against newly emerging threats
Deep-learning-based IDSs	Can handle newly emerging threats	Computationally expensive, requires large amounts of training data
KNN	High detection accuracy (99.99%)	Limited by the quality of the training data
RF	High detection accuracy (99.99%)	Can be computationally expensive
CNN-GRU	High detection accuracy (99.98%)	Requires large amounts of training data

The Table 2 includes information on the approach, the research gap it addresses, and the limitations or drawbacks of the approach.

Table 2. The strengths and limitations of the methods proposed in [3–6].

Approach	Research Gap Addressed	Limitations/Drawbacks
[3]	Improved attack identification	Limited scope and potential for false positives
[4]	Cybersecurity recommender systems	Lack of applications and limited personalization
[5]	Intrusion detection frameworks	Limited focus on IoT devices and potential for false positives
[6]	Beta mixture technique for anomaly detection	Limited action after anomaly detection and potential for false positives

Article [11] highlights the growing threat of industrial sector cyberattacks, which exploit the vulnerabilities of networked machines in the context of Industry 4.0. The rise in investment in innovation and automation has led to a rise in cybersecurity risks, with targeted cyberattacks constantly evolving and improving their attack strategies. These AI-based cyberattacks have the potential to cause exponential damage to organizations. To address this gap, the study analyzes publications of AI-based cyberattacks and derives cybersecurity measures to provide insights for developing defenses against potential future threats.

However, the study also acknowledges the limitations of these measures. For instance, the increasing use of AI in cyberattacks makes it challenging to defend against unknown attacks. Moreover, the interdisciplinary nature of IoT security requires a collaborative approach involving multiple stakeholders, including cybersecurity experts, network architects, system designers, and domain experts [12].

The Table 3 highlights the strengths and limitations of the methods proposed in [11,12]:

Table 3. The strengths and limitations of the methods proposed in [11,12].

Method	Strengths	Limitations
AI-based cyberattack analysis	Provides insights into potential future threats	Limited by the availability of data and the evolving nature of AI-based attacks
Cybersecurity measures derived from AI-based cyberattack analysis	Can be used to make informed decisions regarding cybersecurity measures	May not be effective against unknown attacks
Interdisciplinary approach to IoT security	Recognizes the complexity of IoT security and the need for a collaborative approach	Requires the involvement of multiple stakeholders, which can be challenging to coordinate

In conclusion, the research highlights the importance of proactive measures in addressing the growing threat of industrial sector cyberattacks. However, the limitations of these measures underscore the need for a collaborative approach and ongoing research to stay ahead of evolving threats. The previous table provides a comprehensive overview of the strengths and limitations of the proposed methods, providing valuable insights for future research and cybersecurity strategies.

In [13], the authors discuss the advancements in the connectivity and digitization of critical infrastructure (CI) systems, which have led to enhanced efficiency, productivity, cost savings, and quality. However, these improvements also bring forth the risks associated

with digitalization, such as the generation of more data and increased connectivity. In order to tackle these risks, appropriate CI security solutions must be developed. The paper proposes a novel method to predict cyberattacks by utilizing a proactive approach to identify CI security threats. The foundation of this method is a dataset used to minimize false-positive alerts and to ensure the accuracy of predictions. Machine learning techniques based on real data from various CI sectors are employed to train the dataset and to predict cyberattacks. The prediction mechanism and models depend on factors such as the motivation of the attackers and the nature of the CI. The accuracy of this approach is contingent on the quality of the dataset, which can be improved by incorporating more data. This method can provide valuable insights and information for prioritizing security countermeasures to management and security professionals.

One limitation of the method proposed in [13] is the reliance on the quality of the dataset, which may be difficult to obtain or maintain, especially as the nature of cyberthreats evolves. Additionally, the method may not be adaptable to different CI sectors, as the motivation of the attackers and the nature of the CI can vary significantly between industries.

In [14], the authors discuss the growing threat of cyberattacks on the critical cyberinfrastructure connected to advanced global networks. This infrastructure, which is complex and distributed, generates large amounts of sensitive data and is vulnerable to a range of cyberthreats. The study presents a critical view of the current cybersecurity issues and proposes new approaches, models, and technologies to enhance cybersecurity. The text emphasizes that conventional security protocols are insufficient for addressing the challenges posed by IoT-based CI and suggests that data science and advanced AI techniques will be investigated to develop a more comprehensive and persistent model to deal with massive cyberattacks in the future.

3. Materials and Methods

3.1. Approaches

Machine learning methods used in decision-making algorithms are conventionally divided into two categories [3]:

- Content methods which analyze the types of attacks;
- Collaborative methods which use collaborative filtering.

In content-based filtering, the system recommends items that are similar to the ones a user has liked or interacted with in the past. The system uses features or attributes of the items to determine their similarity, such as text, images, or audio. This method is based on the assumption that, if a user likes one item, they will also like items that are similar to it. Content-based filtering is useful when there is a small number of users and a large number of items, making it difficult to gather collaborative data. It is also useful when the items being recommended have distinct and measurable features, such as books, movies, or songs. Content-based filtering can lead to better accuracy and diversity in recommendations, as it is not influenced by the preferences of other users.

In collaborative filtering, the system recommends items based on the preferences of other users who have tastes and behaviors that are similar to those of the active user. The system uses a collaborative matrix to calculate the similarity between users and items, taking into account the ratings or interactions of all users. This method is based on the assumption that, if a user likes an item, other users with similar preferences will also like it. Collaborative filtering is useful when there is a large number of users and a small number of items, making it easier to gather collaborative data. It is also useful when the items being recommended have subjective features, such as movies, books, or music, where the preferences of other users can provide valuable information. Collaborative filtering can lead to better accuracy and personalization in recommendations, as it takes into account the preferences of other users with tastes similar to those of the active user.

In summary, content-based filtering is more useful when there are distinct and measurable features of the items being recommended, while collaborative filtering is more useful when there are a large number of users and subjective features of the items being

recommended. Both methods have their strengths and weaknesses, and hybrid approaches that combine both methods can often achieve the best results.

Recommender systems can combine both methods to improve their quality. Content-based methods are based on the similarity of element attributes, and collaborative methods calculate the similarity of users' content data based on the matrix of their interactions.

Within a mathematical model of information security, a content-based method can use a set of N attacks and a set of M defense strategies that can be recommended to an information security administrator (Table 4).

Table 4. Content method of the recommender system.

	Attack_1	Attack_2	...	Attack_N
Strategy_1	5	1	...	3
Strategy_2	4	4	...	?
Strategy_3	2	?	...	4
Strategy_4	?	2	...	?
...			...	
Strategy_M	1	?	...	?

In the corresponding cells, there are numerical values denoting the degree of applicability of the defense strategy for a particular attack on a five-point scale, where 5 means Strategy_M is optimal for countering Attack_N and where 1 means Strategy_M is not suitable for protecting against Attack_N.

A learning model for content methods based on retrospective data predicts a particular strategy for each type of attack.

Collaborative methods work with an interaction or rating matrix [4]. The task of machine learning is to determine a function which predicts the importance of an information protection strategy for each attack or for each family of attacks. Such a matrix is usually very large and sparse, with most values missing [5].

In the approach presented here, the base model could not account for zero-day exploits [5], which are understood to be methods used by attackers to attack systems with previously undetected vulnerabilities.

Considering zero-day exploits, the recommender system algorithm could be implemented as follows:

1. Each type of attack/anomaly is represented as a vector of a certain dimensionality.
2. When an attack is detected at the input of a model that is not in the knowledge base, the attack vector is calculated according to step 1.
3. Using the cosine proximity method [6], the cyberattack from the knowledge base that is closest to the newly detected one is selected.
4. A protection strategy is implemented according to the information obtained in step 3.
5. The simplest algorithm calculates the cosine or correlation similarity of rows (users) or columns (elements) and recommends elements that have been selected as KNN [4].

Zero-day exploits pose several challenges to information security:

- **Unknown vulnerabilities:** Zero-day exploits target previously unknown vulnerabilities in software, operating systems, or applications. These vulnerabilities are not yet known to the software vendors or the security community, making it difficult to defend against them.
- **No patches or fixes:** Since the vulnerabilities are unknown, there are no patches or fixes available to address them. This means that organizations and individuals must find alternative solutions to mitigate the risk posed by these exploits.
- **Limited visibility:** Zero-day exploits can be difficult to detect and analyze, as they often use novel techniques and codes that have not been seen before. This limited

visibility makes it challenging to determine the extent of the attack and the damage that has been done.

- High risk: zero-day exploits can be highly risky, as they can be used to gain unauthorized access to systems and data, to steal sensitive information, or to disrupt critical infrastructure.
- Difficult to mitigate: Zero-day exploits can be difficult to mitigate, as they often rely on novel techniques and codes that are not yet understood. This can make it challenging to develop effective defenses against them.
- Targeted attacks: Zero-day exploits are often used in targeted attacks, where the attacker specifically targets a particular organization or individual. These attacks can be highly sophisticated and difficult to detect.

Since zero-day exploits can combine combinations of already known attacks and the recommender system solves the problem of learning without a teacher, with the proposed approach, it would be able to identify the common behavior patterns of a new vulnerability and to show attacks and, consequently, defense strategies that are as similar as possible to the attacks already carried out rather than ignoring them like existing methods to classify a limited set of attacks.

Methods based on matrix factorization [7] reduce the dimensionality of the interaction matrix W of size $n \times v$ and approximate it by two or more small matrices with k latent components, where n is the set of attacks and where v is the set of defense strategies (Figure 1).

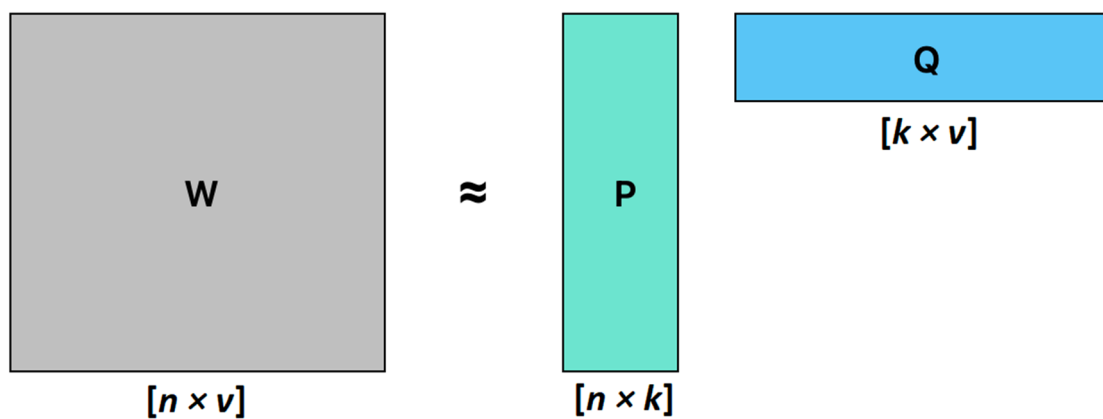


Figure 1. Matrix factorization method.

The matrix factorization method can be described by the following equation [8]:

$$L = \left\| W - P * Q^T \right\|_2 + \lambda (\|P\|_2 + \|Q\|_2), \tag{1}$$

$$r_{ij} = p_i^T q_j = \sum_k p_{ik} q_{kj}, \tag{2}$$

$$\operatorname{argmin}_{q,p} \sum_{i,j} (r_{ij} - p_i^T q_j)^2, \tag{3}$$

where

L is the learning error function of the model;

W is the initial matrix of the interaction between protection strategies and types of attacks;

P and Q are matrices of the rating of the applicability of current information protection strategies for a particular information security event;

$p_{ik} q_{kj}$ are elements of matrices P and Q ;

r_{ij} is the correlation coefficient between the elements of matrices P and Q .

The features of the matrix factorization method [16] can be formulated as follows:

- Each parameter is updated independently;

- The loss error function is calculated with respect to each parameter using the following equation [17]:

$$\begin{cases} p_{ki}^{t+1} = p_{ki}^t + 2\zeta(r_{ij} - p_i^t q_i^t) q_{kj}^t, \\ q_{ki}^{t+1} = q_{ki}^t + 2\zeta(r_{ij} - p_i^t q_i^t) p_{kj}^t, \end{cases} \tag{4}$$

where ζ is the parameter of the protection strategy preference function.

Matrix factorization is a technique commonly used in recommender systems to reduce the dimensionality of large user–item interaction datasets and to identify latent factors that can be used to make personalized recommendations. In the context of information security, matrix factorization can be applied to various problems, such as anomaly detection, intrusion detection, malware detection, risk assessment, and personalized security.

The most popular learning algorithm is the stochastic gradient descent, which minimizes losses through the gradient updating of columns and rows of matrices P and Q , whose error function is described by the following equation [18]:

$$\begin{aligned} \forall p_i : L(p_i) &= \left\| W_i - P_i \times Q^T \right\|_2 + \lambda \|p_i\|_2, \\ \forall q_i : L(q_i) &= \left\| W_i - P_i \times Q_j^T \right\|_2 + \lambda \|q_i\|_2, \end{aligned} \tag{5}$$

where λ is a constant determining the step of parameter change (learning rate). As an alternative to the stochastic gradient descent algorithm, we could use the method of alternating least squares, which interactively optimizes the matrices P and Q [19]. Figure 2 shows a matrix of relationships between different events, where “clumps” are marked with “similar” information security events [20].

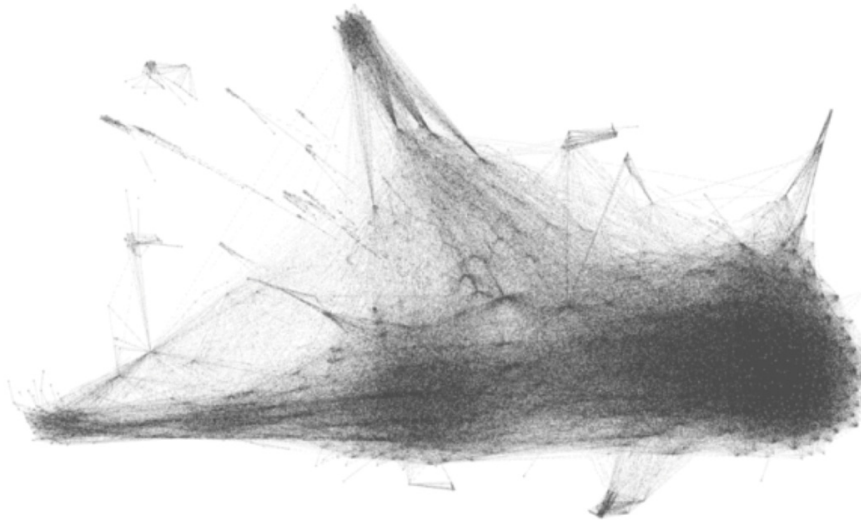


Figure 2. Matrix of associations.

The association matrix rules [21] can be used in a recommendation decision support system. Elements that are often grouped together are connected by an edge in the graph. Let us denote the following:

- I as the set of objects;
- D as the base of transactions;
- S_{\min} as the minimum level of decision support;
- A_{\min} as the minimum confidence threshold.

Rules extracted from the interaction matrix must have at least a minimum level of decision support and a minimum confidence threshold [22] (Figure 3). Support is related to the frequency of occurrence [23]. High confidence means that rules are violated infrequently.

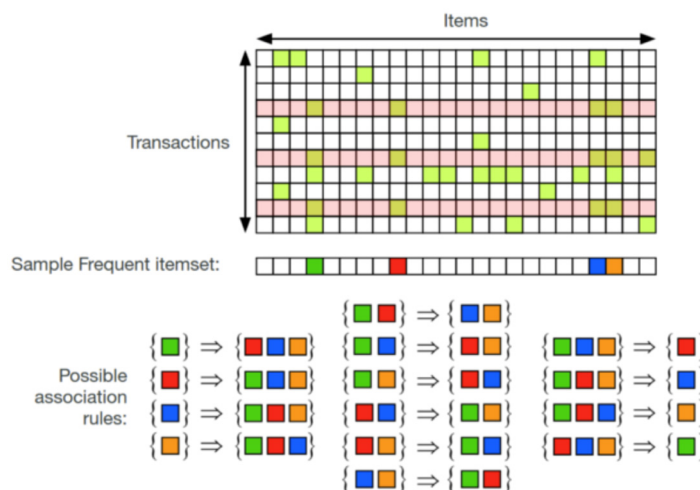


Figure 3. The APRIORI algorithm.

The scalability of the rule search can be improved with the APRIORI algorithm [19] (Figure 3), which examines the state space of possible sets of frequent items and removes branches of the search space that are not frequent [24].

The main metric for recommendation quality is the normalized discounted cumulative gain (nDCG) metric [25]. The advantage of this metric is its finiteness: it takes values in the range [0;1]. The closer its value is to 1, the better the ranking of protection strategies for a particular event is [26]:

$$nDCG@K = \frac{DCG@K}{IDCG@k'} \tag{6}$$

where

$$DCG@K = \sum_{k=1}^K \frac{2^{r^{true}(\pi^{-1}(k))} - 1}{\log_2(k + 1)} \tag{7}$$

takes into account the order of items in the list by multiplying the relevance of an item by a weight equal to the inverse logarithm of the item number [27], IDCG@K is the ideal value of DCG metric, and k is the order number in the ranked list.

The normalized discounted cumulative gain (DCG) metric is a metric that is widely used to evaluate the quality of recommendations in information retrieval and recommendation systems. It measures the usefulness or gain of a recommendation based on the user’s feedback, such as clicks or purchases. The DCG metric is calculated as the sum of the gains of all recommendations, where the gain of each recommendation is discounted by a decaying function of the time elapsed since the recommendation was made. The discounting factor allows the metric to prioritize more recent recommendations, as they are considered to be more relevant and useful to the user.

The DCG metric is normalized to ensure that the scores are on the same scale regardless of the number of recommendations made. Normalization is typically done by dividing the DCG score by the maximum possible score, which is the sum of the gains of all recommendations. To assess the quality of recommendations using the DCG metric, we could compare the score of each recommendation to a threshold value. If the score is above the threshold, the recommendation is considered to be of high quality and is likely to be relevant to the user. If the score is below the threshold, the recommendation is considered to be of low quality and may not be relevant to the user.

Thus, a decision support system for information security tasks can be built using two methods [28] based on the following:

- Collaborative filtering of information security events;
- Matrix decomposition of the matrix “cyberattack–strategy”.

Both methods assume that the knowledge base contains retrospective data about the administrator's actions when an information security event occurs [29]. However, the collaborative filtering method is resistant to zero-day vulnerabilities.

3.2. Proposed Recommender System

A general algorithm for conducting research on the construction of a recommender system in the field of IS is as follows:

1. Review the basic algorithms for building recommender systems, which are widely used at this moment in all areas of the information industry.
2. Analyze and collect a set of data for the research. A prerequisite is that the dataset must relate specifically to the specifics of anomaly detection or attack classification in order to realize a decision support system based on it in the field of information security.
3. Determine the hypothesis and possible variants of the research. Identify the algorithms that can be used to solve the problem. A hypothesis about the possibility of transferring knowledge about the attacks from the vector space of the features into semantic form using semantic similarity (cosine similarity and Pearson correlations or normalized cosine similarity [30]) was formulated.
4. Formulate metrics for evaluating the quality of the developed algorithm.
5. Evaluate the approaches under consideration with respect to maximizing the quality metric.
6. Describe the advantages and disadvantages of the chosen approach.

4. Experiment

This section identifies the datasets used in this study, the exploratory dataset, the explanation and reason for the applicability of the quality metrics, and the progress of the experiments.

4.1. Exploratory Data Analyses

This study examined the UNSW-NB 15 dataset. Unprocessed network packets from the UNSW-NB 15 dataset were created using the IXIA PerfectStorm software in the laboratory of the Australian Cyber Security Centre (UNSW Canberra) to create a set of normal real-world actions and synthetic modern attacks. This dataset allows for simultaneous training to identify normal system behavior and different types of anomalies.

Using the utility "tcpdump" [31], which captures and analyzes the network traffic, 100 GB of raw traffic was generated. Figure 4 shows the general data collection algorithm used during the simulation for the implementation of attacks:

- Pcap files were generated using the tcpdump tool;
- The signs of network traffic were extracted with the Argus and Bgo IDS tools;
- Synthetic traces were generated, and records were stored in a database.

This dataset includes data on abnormal system behavior when exposed to nine types of attacks:

- Fuzzer;
- Analysis;
- Backdoors;
- DoS;
- Worm;
- Shellcode;
- Reconnaissance;
- Generic;
- Exploit.

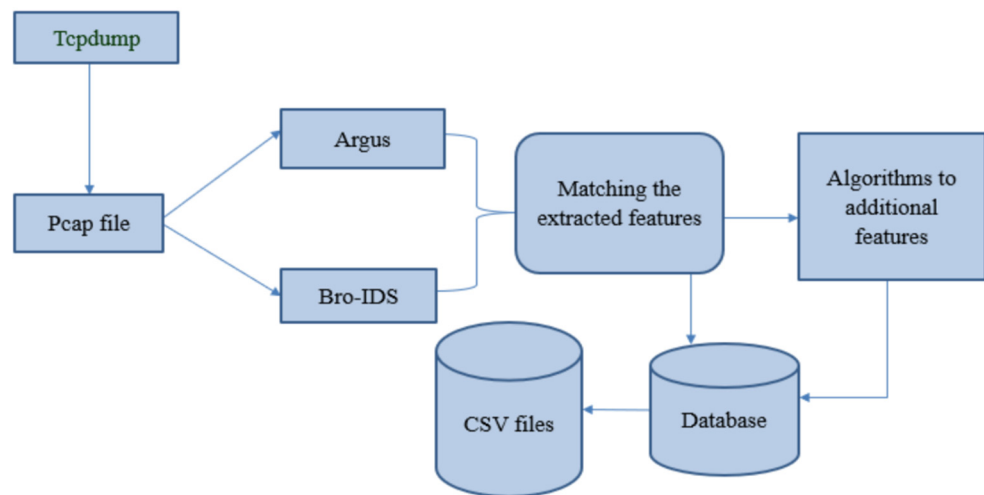


Figure 4. General scheme of data collection.

Figure 5 shows the distribution of the target variable, which is close to exponential in shape. We can clearly see that the sample is unbalanced and contains the largest number of records with normal system behavior as well as data affected by the generic, exploit, and fuzzer attacks. The smallest number of records contain data affected by backdoor, shellcode, and worm attacks. Each event is a set of 42 attributes of the system state, which is represented in the form of a vector of finite dimensionality.

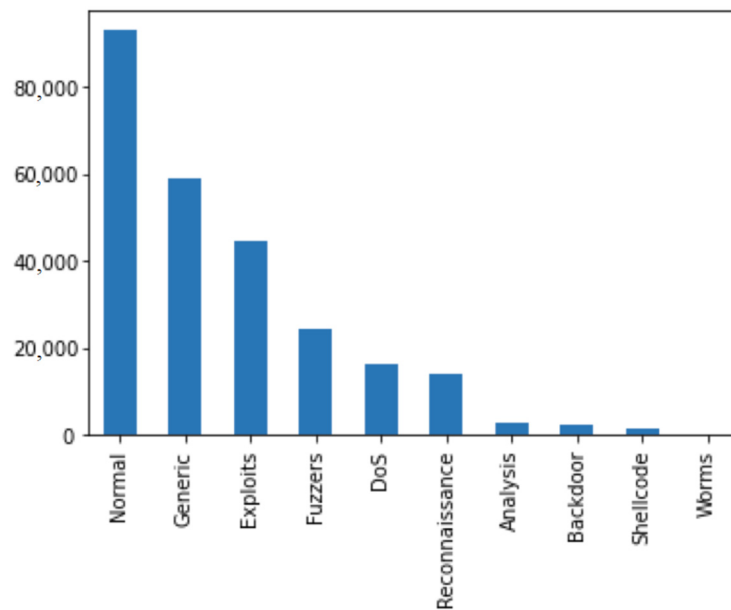


Figure 5. Distribution of the target variable by attack type.

A hypothesis about the possibility of transferring knowledge about the attacks from the vector space of the features into a semantic form using semantic similarity (cosine similarity and Pearson correlations or normalized cosine similarity [31]) was formulated.

4.2. Semantic Proximity Methods

Let us consider each of these methods in more detail:

1. When cosine similarity is used, the closeness of two attacks is calculated as the cosine of the angle between the vectors corresponding to their rows in the score matrix [32]. Thus, the cosine similarity of users u and v is defined by the following:

$$sim(u, v) = \frac{\sum_{i \in I} r_{ui} r_{vi}}{\sqrt{\sum_{i \in I} r_{u,i}^2} \sqrt{\sum_{i \in I} r_{v,i}^2}}, \tag{8}$$

where r_{ui} is the vector element u and r_{vi} is the vector element v .

2. The Pearson correlation coefficient reflects the degree of linear dependence between two centered vectors. The closeness is determined by the extent to which the system parameters for the two time sections are similar to each other [33]. For the user vectors u and v , the correlation coefficient Formula (2) takes the following form:

$$sim(u, v) = \frac{\sum_{i \in I} (r_{ui} - \bar{r}_u)(r_{vi} - \bar{r}_v)}{\sqrt{\sum_{i \in I} (r_{ui} - \bar{r}_u)^2} \sqrt{\sum_{i \in I} (r_{vi} - \bar{r}_v)^2}} \tag{9}$$

where r_{ui} is the vector element u and r_{vi} is the vector element v .

3. The normalized cosine similarity [34] computes the user similarity like the cosine convergence but does so using the vectors of the deviation in the user ratings from the average object ratings [35]. Thus, the more similar the user ratings for some object, the less deviation there is from the “generally accepted” ratings of the object, and the more similarity the function shows between users [36].

The normalized cosine similarity is a measure of the similarity between two vectors that is often used in the context of document similarity in information retrieval or text classification tasks. Given two normalized vectors (i.e., vectors with a length of one) A and B, the cosine of the angle between them is computed. This cosine value ranges from -1 to 1 , where -1 indicates complete orthogonality (no similarity), 0 indicates orthogonality (no similarity), and 1 indicates complete similarity. The normalized cosine similarity is obtained by dividing the cosine value by its maximum possible value, which is one, resulting in a value between zero and one. This normalization ensures that the similarity values are comparable across different vector lengths and dimensions.

Pearson correlation is a measure of the linear relationship between two continuous variables. It is based on the covariance of the variables, normalized by their standard deviations, resulting in a value between -1 and 1 . A value of 1 indicates a perfect positive linear relationship; 0 indicates no linear relationship; and -1 indicates a perfect negative linear relationship. Pearson correlation is particularly useful when the variables are measured on an interval or ratio scale and they follow a linear trend. It is widely used in various fields, including finance, psychology, and social sciences, to assess the strength and direction of the linear relationship between two variables.

After selecting a similarity function for each attack, it is necessary to determine the set of close attacks K , the estimates of which will add up to the estimate of the object. To do this, the following approaches are used [37]:

- Setting a threshold: the user whose proximity measure exceeds a certain value is considered a neighbor.
- Finding the KNN: the set consists of k users with the greatest similarity, where k is a preselected constant.

The choice between one or another approach is determined by what is more important for the calculation of the estimate, quality (the first approach) or quantity (the second approach) [38]. By having a set of close users, we can find the estimate using Formula (4):

$$\widehat{r}_{ui} = \bar{r}_u + \frac{\sum_{i \in I} (r_{ui} - \bar{r}_u) sim(u, k)}{\sum_{i \in I} |sim(u, k)|}, \tag{10}$$

where $sim(u, k)$ are the cosine similarity vectors u and k , \bar{r}_u is the mean vector u , and \bar{r}_k is the mean vector k .

In this study, the events were grouped into clusters using the methods described above. Each event was represented as a vector of dimension 86. A quality assessment

was measured using the heap, data homogeneity, and elbow methods. The most sensitive metrics were the heap and data homogeneity methods.

4.3. The Reliability of the Proposed System

Figure 6 shows the distribution of the five states of the system on a two-dimensional plane for easier visualization, which was compressed using the principal component method. The graph indicated below was plotted using the following algorithm: each event was compared to a known set of one-to-all attacks, the cosine of the proximity was calculated, and the given event referred to the type of attack that was closest [29]. The different states of the system are highlighted with different colors: blue—normal, light green—generic, yellow—exploits, purple—fuzzers, and green—DoS.

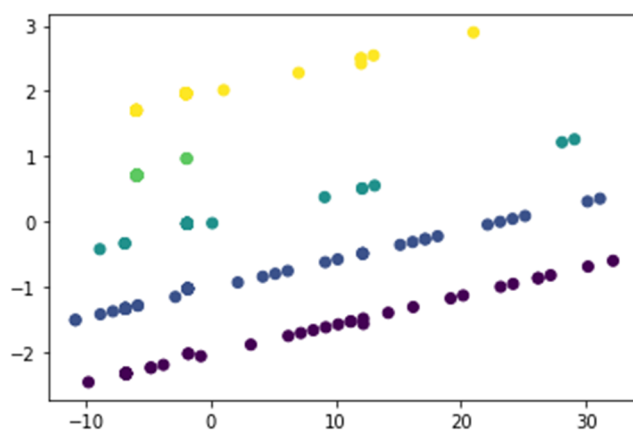


Figure 6. Distribution of attack types on the two-dimensional plane (different colors are different types of attacks).

Table 5 compares the three approaches with respect to quality metrics such as uniformity and fidelity. The method based on the cosine closeness of two vectors had the highest quality metrics. Pearson correlation showed the worst homogeneity of the data in one cluster. Data homogeneity was defined as the percentage of identical attack families in the same cluster.

Table 5. Comparison of quality metrics.

	Cosine Similarity	Pearson Correlation	Normalized Cosine Similarity
Homogeneity	76%	64%	72%
Coverage	68%	44%	60%

The advantage of this approach is the independence of determining the similarity between attack types from the distance between the points in the Euclidean space since this distance does not determine the angle between the corresponding vectors: the smaller the angle, the higher the similarity is.

As part of the approach evaluation, the LAST.FM dataset was analyzed to build the recommender system. The authors of [15] used it to build a hybrid recommender system, named NRH (Node2vec-side and Ripplet Hybrid Model), for personalized recommendations. This dataset includes 92,834 rating data from 1892 users and 17,632 implementers, and the corresponding knowledge graph contains 15,518 triples with 9366 entities with 60 links. For each user, a vector of their behavior was generated and projected into the multidimensional space. Table 6 shows the results of the collaborative filtering approaches and the previously proposed approach. This approach showed a small increase in coverage (by 1%) and a marked improvement in homogeneity (by 15%).

Table 6. Comparison of collaborative filtering and an event recommendation approach using cosine proximity.

	Collaborative Filtering	An Event Recommendation Approach Using Cosine Proximity
Homogeneity	64%	79%
Coverage	83%	84%

5. Discussion

As part of this study, the main algorithms of recommendation systems were considered. The following problems of modern recommendation systems were highlighted [39]:

- The insufficient accuracy of recommendations: A lack of data on users or items can lead to problems with the accuracy of recommendations. If the system does not have enough information on user preferences or item descriptions, it may give incorrect recommendations or miss relevant suggestions. For example, if the user does not have a large volume of interactions with the system or if the subjects have limited data, then the recommendations may not be too accurate.
- The problem of the “filtering bubble”: Recommendation systems can create “filtering bubbles” when recommendations are limited to the preferences and interests of the user. This means that users may be limited in their experience since the system offers them only those items or content that match their previous preferences. As a result, users may miss out on new and diverse offers that might interest them.
- The cold start problem: When a recommendation system encounters new users or new items, it may experience the cold start problem. This happens when the system does not have enough information to create relevant recommendations. This makes it difficult to create accurate recommendations in such situations.
- The problem of a lack of diversity: recommendation systems sometimes tend to offer items or content that are too similar based on the user’s previous preferences.
- The interpretation of the results: recommendation systems often do not provide explicit explanations or justifications for their recommendations, which can cause distrust and dissatisfaction among users.

As mentioned above, most researchers implement various machine learning models for the early detection of attacks and anomalies, but it is worth paying special attention to the cases of what to do after a specific attack has been detected and what actions should be performed. In this study, an approach to building a recommendation system in the field of information security was proposed rather than an algorithm—a method to detect an attack. A successful experiment was conducted to build such a system on a specific dataset. The advantage of the proposed approach is the independence of determining the similarity between attack types from the distance between the points in the Euclidean space since this distance does not determine the angle between the corresponding vectors: the smaller the angle, the higher the similarity is.

The proposed recommender system can be used in a production environment in the following scenarios:

- (1) The configured logging system automatically generates an algorithm for the information security administrator’s actions when an anomaly/attack is detected in the network and, subsequently, when a similar event is detected; it indicates in advance that event N is similar to a given percentage of an earlier event K , where the given actions were taken.
- (2) The security administrator, while investigating the incident, automatically records the conclusions on anomaly prevention, and, when a new, very similar event occurs, this algorithm is immediately shown to the administrator.

There are two limitations:

- The first is either setting up a good logging system or manually marking, by the information security administrator, the implemented protection strategies, forming a so-called knowledge base so that the system can be guided by this knowledge base when recommending future protection strategies.
- The second is the need for a knowledge base, as the problem of a cold start is acute.

In the future, it is proposed to develop approaches to create a unified base of information security strategies in order to level out the limitations identified earlier.

One of the disadvantages of this method is that it requires the periodic verification by the security administrator of past anomalies to reduce false positives.

However, for a new anomaly/attack, which may combine several previously conducted attacks, this algorithm will produce a set of defense strategies for all the events that are similar to the current one by a given threshold, which increases the stability of the system. In addition, this approach allows the semiautomated marking of events.

6. Conclusions

In this study, the approaches to building recommender systems in information security were considered. Each approach has its own advantages and disadvantages, but an important factor in the application of collaborative filtering is its resistance to zero-day vulnerabilities. The study hypothesized that, by using the semantic proximity of different types of attacks, it is possible to implement an advisory model to search, rank, and issue protection strategies which have been applied to similar attacks. For this purpose, the applicability of the cosine similarity, Pearson correlation, and normalized cosine similarity methods in building recommender systems in information security was analyzed. The most qualitative partitioning of attack families was obtained using the cosine similarity method, with the quality metric being the proportion of homogeneous objects in the same cluster. Thus, a qualitative partitioning of attack types into clusters based on their semantic state was obtained, which allowed us to apply the cosine similarity method to determine the type of attack.

Author Contributions: Conceptualization, A.S.; methodology, A.S. and A.R.; software, G.A.; validation, G.A., A.R. and A.S.; formal analysis, A.R.; investigation, G.A.; resources, G.A.; data curation, G.A.; writing—original draft preparation, G.A.; writing—review and editing, A.S. and A.R.; visualization, G.A.; supervision, A.S.; project administration, A.S.; funding acquisition, G.A., A.S. and A.R. All authors have read and agreed to the published version of the manuscript.

Funding: The research was funded by the Russian Science Foundation (project No. 22-71-10095).

Data Availability Statement: The data and source code presented in this study are available on request from the corresponding author.

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

References

1. Griffiths, C. The Latest 2023 Cyber Crime Statistics. Available online: <https://aag-it.com/the-latest-cyber-crime-statistics/> (accessed on 11 June 2023).
2. Frackiewicz, M. The Role of Artificial Intelligence in Cybersecurity Threat Detection, Artificial Intelligence, TS2 Spaceon. 18 June 2023. Available online: <https://ts2.space/en/the-role-of-artificial-intelligence-in-cybersecurity-threat-detection/> (accessed on 11 June 2023).
3. Bolshev, A.K. *Algorithms of Traffic Transformation and Classification for Intrusion Detection in Computer Networks*; Abstracts of V.I. Ulyanov (Lenin) LETI; Saint-Petersburg State Electrotechnical University (SPbGETU): St. Petersburg, Russia, 2011; pp. 134–151.
4. Vitenburg, E.A. Formalized model of intellectual decision support system in the field of information protection. Proceedings of TulSU. Technical Sciences. 2017. No. 7.
5. Abdullahi, M.; Baashar, Y.; Alhussian, H.; Alwadain, A.; Aziz, N.; Capretz, L.F.; Abdulkadir, S.J. Detecting Cybersecurity Attacks in Internet of Things Using Artificial Intelligence Methods: A Systematic Literature Review. *Electronics* **2022**, *11*, 198. [CrossRef]
6. Grigaliūnas, Š.; Brūzgienė, R.; Venčkauskas, A. The Method for Identifying the Scope of Cyberattack Stages in Relation to Their Impact on Cyber-Sustainability Control over a System. *Electronics* **2023**, *12*, 591. [CrossRef]

7. Smirnov, A.A.; Salyp, B.Y. Analysis of software models to determine the measure of semantic proximity of natural language sentences. *Student* **2022**, *5*, 3498–3508.
8. Moustafa, N.; Creech, G.; Slay, J. Anomaly Detection System Using Beta Mixture Models and Outlier Detection. In *Progress in Computing, Analytics and Networking; Advances in Intelligent Systems and Computing*; Springer: Singapore, 2018; Volume 710. [[CrossRef](#)]
9. Mazhar, T.; Irfan, H.M.; Khan, S.; Haq, I.; Ullah, I.; Iqbal, M.; Hamam, H. Analysis of Cyber Security Attacks and Its Solutions for the Smart grid Using Machine Learning and Blockchain Methods. *Future Internet* **2023**, *15*, 83. [[CrossRef](#)]
10. Alzahrani, A.; Aldhyani, T.H.H. Design of Efficient Based Artificial Intelligence Approaches for Sustainable of Cyber Security in Smart Industrial Control System. *Sustainability* **2023**, *15*, 8076. [[CrossRef](#)]
11. de Azambuja, A.J.G.; Plesker, C.; Schützer, K.; Anderl, R.; Schleich, B.; Almeida, V.R. Artificial Intelligence-Based Cyber Security in the Context of Industry 4.0—A Survey. *Electronics* **2023**, *12*, 1920. [[CrossRef](#)]
12. Tariq, U.; Ahmed, I.; Bashir, A.K.; Shaukat, K. A Critical Cybersecurity Analysis and Future Research Directions for the Internet of Things: A Comprehensive Review. *Sensors* **2023**, *23*, 4117. [[CrossRef](#)] [[PubMed](#)]
13. Alqudhaibi, A.; Albarrak, M.; Aloheel, A.; Jagtap, S.; Salonitis, K. Predicting Cybersecurity Threats in Critical Infrastructure for Industry 4.0: A Proactive Approach Based on Attacker Motivations. *Sensors* **2023**, *23*, 4539. [[CrossRef](#)]
14. Djenna, A.; Harous, S.; Saidouni, D.E. Internet of Things Meet Internet of Threats: New Concern Cyber Security Issues of Critical Cyber Infrastructure. *Appl. Sci.* **2021**, *11*, 4580. [[CrossRef](#)]
15. Ni, W.; Du, Y.; Ma, X.; Lv, H. Research on Hybrid Recommendation Model for Personalized Recommendation Scenarios. *Appl. Sci.* **2023**, *13*, 7903. [[CrossRef](#)]
16. Chertov, O.; Brun, A.; Boyer, A.; Aleksandrova, M. Comparative analysis of neighborhood-based approach and matrix factorization in Recommender systems. *East.-Eur. J. Enterp. Technol.* **2015**, *3*, 4–9.
17. Zhang, Z.; Afanasiev, G.I. Basic technologies and prospects for the evolution of personalized recommender systems. *E-SCIO* **2022**, *4*, 309–320.
18. Razuvaev, K.A.; Grinberg, H.E.; Maslova, A.S.; Veinskii, V.A.; Milutin, A.B. Analysis of modern approaches in the design of recommendation systems. *Int. J. Appl. Sci. Technol. Integral* **2021**, *2*, 253–261.
19. Pavlov, P.S. Methods for assessing the quality of recommendation systems. *Int. J. Humanit. Nat. Sci.* **2018**, *6*, 178–182.
20. Smolenchuk, T.V. Collaborative filtering method for recommendation services. *Bull. Sci. Educ.* **2019**, 18–21.
21. Smirnov, V.M.; Matveev, S.P. Methods of protection against malicious software. *StudNet* **2022**, *5*, 1595–1599.
22. Fisun, V.V. *Artificial Intelligence of Information Security Management of Critical Information Infrastructure Objects: A Monograph*; Rusayns: Moscow, Russia, 2022; pp. 17–21.
23. Fisun, V.V. Methodology of security assessment in the intellectual system of information security management of critical information infrastructure objects. *NAU* **2018**, 2–10.
24. Kalandarov, I. Assessment of Information Security Risks in Ensuring the Confidentiality of Information Resources. *Probl. Comput. Sci. Energy* **2017**, *6*, 42–48.
25. Kurinnikh, D.Y.; Aidinyan, A.R.; Tsvetkova, O.L. Approach to the clustering of threats to information security of enterprises. *IVD* **2018**, 91.
26. Aidinyan, A.R.; Tsvetkova, O.L.; Kikot, I.R.; Kazantsev, A.V.; Kaplun, V.V. On the approach to assessing the information security of an enterprise. In *Proceedings of the System Analysis, Management and Information Processing: Collected Works of the V International Scientific Seminar, Divnomorskoye Settlement, Tuapse, Russia, 2–6 October 2014*; pp. 109–111.
27. Tsvetkova, O.L.; Zaslono, S.A. Simulation modeling of the dependence of information security of the organization on the field of activity. *DSTU Bull.* **2017**, 116–121.
28. Tsvetkova, O.L.; Aidinyan, A.R. Intellectual system of information security assessment of the enterprise from internal threats. *Bull. Comput. Inf. Technol.* **2014**, 48–53.
29. Kozunova, S.S.; Kravets, A.G. Formalized Description of Information System Risk Management Procedure. *Vestn. Astrakhan State Tech. Univ. (Ser. Manag. Comput. Sci. Inform.)* **2018**, *2*, 61–70.
30. Tyurin, A.G.; Zuev, I.O. Cluster analysis, methods and algorithms of clustering. *Russ. Technol. J.* **2014**, *2*, 86–97.
31. Mahruse, N. Modern trends in data mining methods: The method of clustering. *Mosc. Econ. J.* **2019**, 359–377.
32. Kadar, C.; Maculan, R.; Feuerriegel, S. Publicdecision support for low population density areas: An imbalance-aware hyperensemble for spatio-temporal crime prediction. *Decis. Support Syst.* **2019**, 107–117. [[CrossRef](#)]
33. Rzaev, R.R. Information system to support procedural decision making. *Syst. Means Inform.* **2016**, 182–198.
34. Duga, S.; Sebyakin, A.; Nosyreva, L. The concept of a decision support system in the preliminary investigation. *Inf. Technol. Secur.* **2019**, *26*, 45–57. [[CrossRef](#)]
35. Duga, S.; Trufanov, A. The knowledge graph concept of decision support system in preliminary investigation. *Secur. Inf. Technol.* **2020**, *22*, 55–66. [[CrossRef](#)]
36. Tushkanova, O.N.; Samoilo, V.V. KnowledgeNet: A model and system of accumulation, representation and use of knowledge and data. *Des. Ontol.* **2019**, *9*, 117–131. [[CrossRef](#)]
37. Podruzhkina, T.A.; Fedorov, D.Y. Algorithms for Planning the Learning Process on the Basis of Semantic Knowledge Networks. *Bull. St.-Petersburg Univ. State Fire Serv. EMERCOM Russ.* **2017**, *2*, 107–116.

38. Vasiliev, V.I.; Belkov, N.V. Decision support system for the security of personal data. *Bull. UGATU* **2011**, 45–52.
39. Balraj, K.; Neeraj, S. Approaches, Issues and Challenges in Recommender Systems: A Systematic Review. *Indian J. Sci. Technol.* **2016**, 9. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.