A cross-domain access control mechanism based on model migration and semantic reasoning

Ming Tan[1,2], Aodi Liu[1,2]*, Xiaohan Wang[1,2], Siyuan Shang[1,2], Na Wang[1,2], and Xuehui Du[1,2]
[1] Information Engineering University
Zhengzhou 450001, Henan – P. R. China
[2] He'nan Province Key Laboratory of Information Security
Zhengzhou 450001, Henan – P. R. China
[e-mail: ladyexue@163.com, ladyexue@163.com, wang523648@163.com, a525435400@163.com, twftina_w@126.com, dxh37139@163.com]
*Corresponding author: Aodi Liu

Received November 12, 2023; revised January 8, 2024; accepted June 3, 2024;
published June 30, 2024

Abstract

Access control has always been one of the effective methods to protect data security. However, in new computing environments such as big data, data resources have the characteristics of distributed cross-domain sharing, massive and dynamic. Traditional access control mechanisms are difficult to meet the security needs. This paper proposes CACM-MMSR to solve distributed cross-domain access control problem for massive resources. The method uses blockchain and smart contracts as a link between different security domains. A permission decision model migration method based on access control logs is designed. It can realize the migration of historical policy to solve the problems of access control heterogeneity among different security domains and the updating of the old and new policies in the same security domain. Meanwhile, a semantic reasoning-based permission decision method for unstructured text data is designed. It can achieve a flexible permission decision by similarity thresholding. Experimental results show that the proposed method can reduce the decision time cost of distributed access control to less than 28.7% of a single node. The permission decision model migration method has a high decision accuracy of 97.4%. The semantic reasoning-based permission decision method is optimal to other reference methods in vectorization and index time cost.

Keywords: Cross-domain, access control, model migration, semantic reasoning, blockchain, data security

https://doi.org/10.3837/tiis.2024.06.010 **ISSN : 1976-7277**

1. Introduction

With the widespread popularization of network and information technology, a complex network environment with many characteristics such as openness, mobility, heterogeneity, and coexistence of multiple security domains has gradually formed, and there are a large number of security domains with independent databases and information systems in the complex network environment. At the same time, due to the rapid development of new computing paradigms such as big data and the Internet of Things, the efficiency of data sharing and utilization has been greatly improved, and the sharing of data resources between different security domains has become the norm, which creates great value. However, while the utilization of data brings great value, the problem of data security is also becoming more and more prominent. As one of the effective security mechanisms to protect data sharing, access control mechanisms are facing great challenges in the new computing environment.

First of all, in the process of cross-domain circulation and sharing of data in the new computing environment, in the face of distributed and highly concurrent access control requests, the centralized access control mechanism faces the performance bottleneck problem, while the distributed access control mechanism faces the security and trustworthiness problem of permission information.

Secondly, in a distributed environment, different security domains have independent and heterogeneous access control systems, and the access control policies formulated for data resources are also different. On the one hand, when users between heterogeneous systems make cross-domain access requests, they need to negotiate policies between heterogeneous systems, which leads to the difficulty of policy negotiation and inefficiency of cross-domain access control; on the other hand, the traditional access control mechanisms in security domains have different access control languages, which leads to the inefficiency of cross-domain access control. The new mechanism needs to reformulate the access control policy, and the new policy needs to take into account the permissions granted by the old policy, which adds a huge amount of workload for the policy makers and improves the security risk of the access control policy. Therefore, it is difficult to get both efficient and secure consensus on access control privileges for cross-domain access by users between heterogeneous systems.

Finally, in the real world of big data, unstructured data accounts for 75% to 80% of the data. The dynamic, distributed, and multi-source heterogeneous characteristics of massive unstructured data resources raise the following needs for access control: the need to accurately describe unstructured data resources at a fine-grained level, the need to formulate a cross-domain access control policy in combination with the security intent of complex environments, and the need to achieve efficient permission decisions for massive and dynamic unstructured resources.

The contributions of this research can be summed up as follows:

- For the first time, CACM-MMSR (Cross-domain Access Control Mechanism Based On Model Migration And Semantic Reasoning) is proposed to provide a trusted platform for the access control functions of different security domains in a distributed environment by relying on smart contracts in blockchain. The smart contract serves as a carrier for the interactive transmission of permission information and access control decision function within the access control mechanism, realizing the automation and systematization of access control decision in distributed environments.
- A permission decision model migration method based on access control logs is designed. We use ResNet network to learn the access control log information in a

security domain and generate a migration model, which contains the access control decision rules of the historical policies in the security domain, so as to "reverse inference" the historical policy information in the form of a model. It realize the migration and replacement of the historical policies. This method uses the ResNet model to unify the access control language of each security domain, which solves the problem of difficult policy negotiation between heterogeneous security domains, improves the efficiency of cross-domain access control, and solves the problem of compatibility between new and old policies.

- A semantic reasoning-based permission decision method for unstructured data is designed. The unstructured text is transformed into semantic vectors through semantic models, that is, the unstructured text is described at a fine-grained way at the semantic level. Then using similar text retrieval technology, the candidate set with the highest similarity (Top1) is quickly reasoned out among the massive permissions text, and the threshold decision is made by comparing the similarity value of the candidate set with the value of the similarity threshold, to realize efficient permissions decision on massive dynamic unstructured texts.

The rest of the paper is organized as follows: section II discusses related work; section III gives an overview of the CACM-MMSR framework; section IV and V detail the principles of the access control log-based permission decision model migration method and the semantic reasoning-based permission decision method for unstructured text, respectively. Section VI analyzes the performance of the two methods through experimental comparison, proving the feasibility of the two methods in the new computing environment. Section VII summarizes the whole paper.

2. Related Work

Access control technology guarantees that resources can only be used by legitimate users to execute legal operations according to pre-set access control policies, preventing illegal authorized access to resources. With the emergence of new computing environments such as big data and Internet of Things, the cross-domain access control for new computing environments has been one of the research directions for researchers in the field of computer security.

According to the U.S. Department of Defense's Trusted Computer System Evaluation Criteria (TCSEC), access control can be categorized into Discretionary Access Control (DAC) and Mandatory Access Control (MAC), and in the process of the development of Internet applications, Role-Based Access Control (RBAC) has emerged. However, the traditional closed-environment access control models, such as DAC, MAC, RBAC, etc., are facing the challenge of exploding permission relationships in new computing environments. Attribute-Based Access Control (ABAC) effectively solves the problem of fine-grained access control in new computing environments characterized by large-scale, strong dynamics, and strong privacy by using attributes as the key elements of access control, and is the mainstream model studied by researchers in new environments. Fang[1] et al. analyze the key issues facing the overall process of attribute-based access control, the current state of research and the development trend, and point out the future research direction of attribute-based access control. Firstly, the dynamic, distributed, and multi-source heterogeneity of unstructured data leads to difficulties in extracting the internal attributes of the data with the permission information in the ABAC model.

To solve this problem, Wu[2] et al. mined ABAC policies that fit the subject's behavioral patterns from access logs and attribute, and analyzed the correctness and semantic quality of the policies to get the semantic-rich ABAC policy set, which provides a powerful support for security administrators to build, maintain, and optimize policies Aiming at the problem of automated generation of access control policies, Liu[3] et al. proposed a deep learning-based ABAC access control policy generation framework to extract attribute-based access control policies from natural language text, which can significantly reduce the time cost of access control policy generation and provide effective support for the implementation of access control. In addition, in complex application scenes for cross-domain access control, researchers have tried to combine ABAC with new technologies to better adapt to the access control requirements in new environments. For example, Zhang[4] et al. proposed a blockchain-based inter-domain access control model that combines the ABAC model and blockchain to provide standardized secure, convenient, autonomous and fine-grained access control for inter-domain access. However, with the emergence of different application scenes in new computing environments, the traditional ABAC model still fails to meet the respective access control requirements, and thus scene-specific access control models emerge.

Semantic Attribute-Based Access Control (SABAC) refers to the use of semantic information in access control to restrict access to resources, and by combining ABAC with semantic techniques to consider the semantics of attributes, it not only facilitates interoperability, but also enhances the expressiveness of the access control policy to match entity attributes with the attributes (syntactically) used in the access control policy in distributed and heterogeneous environments. Hamed[5] et al. comprehensively reviewed the research efforts in the development of SABAC and identified opening questions and possible research by showing the strengths and weaknesses of previous research.

Trust-based access control is a security mechanism to control access permissions based on the trust level of a user, device, or other entity, which in turn improves the security of access control to sensitive resources. Khan[6] et al. proposed a trust-based access control mechanism for data security on cloud platforms, where before granting access to a user, user behaviors, network behaviors, demand behaviors, and security behaviors were analyzed data to calculate the trust value and grant or deny access control based on the trust value result.

Rights-based access control is a security mechanism for controlling access rights based on specific rights such as a user's position or the location of the organizational structure to which the user belongs, which can help organizations to better meet the security needs for different positions. Sabrina[7] et al. proposed a rights-based blockchain access control architecture for real-world IoT environments in the context of application domains for large-scale IoT deployments that helps resource owners to securely delegate access to any entity outside the organizational trust boundaries.

Risk-based access control refers to the process of adaptive authorization of access requests by weighing the security level and the risk factors of access requirements. Ma[8] et al. proposed a paragraph-level authorization text data access control model based on risk-aware topics (RTBAC), in order to solve the problem of unstructured textual data that is difficult to extract the internal attributes in the ABAC model and thus is not effectively authorized. Using topics to represent the content relationship between users and data, thus fine-grained access control services are performed for unstructured text. Risk-based access control can flexibly handle the content relationship between the user and the data.

However, the above access control technology faces the following problems in the new computing environment. First, there is a lack of integration with distributed processing architectures, which results in lower performance when processing large amounts of data and can even lead to a single point of failure. Second, different security domains have different access control mechanisms and different policy languages in the scene of access control heterogeneity. Access control mechanisms of different security domains are hardly compatible with each other, and cross-domain policy negotiation is difficult in cross-domain access control. Third, when the traditional access control model is updated to the new one, it is difficult to retain the historical policy information completely because of the semantic difference between the old and new access control policies. And with the expansion of the scale of the policy, the burden of access control model engineering and updating is difficult to retain. Finally, the granularity of access control is also difficult to achieve at the semantic level. We design CACM-MMSR, which can effectively resolve the above problems and provides new ideas for cross-domain access control.

3. CACM-MMSR framework

The CACM-MMSR framework realizes the interactive transmission of access control information and the automated execution of access control decision functions in a distributed cross-domain background based on smart contracts, which mainly consists of two modules and is realized based on two methods respectively. The access control log-based permission decision model migration method uses the ResNet[9] neural network to learn the historical policy information contained in the logs, so as to inverse inference the historical access control policies. Using this access control log migration model for decision making avoids leakage of access control log information, improves the efficiency of decision making, and realizes the compatibility of old and new access control models; meanwhile, using a unified model language to represent the historical policy information, it solves the problem of policy language inconsistency caused by the heterogeneity of old and new access control mechanisms. The similarity decision threshold is set in combination with practical application requirements, so as to flexibly and fine-grained decide the authority of unstructured text based on semantic reasoning.

The overall procedure of CACM-MMSR is shown in **Fig. 1**. In the preparation step of permission decision model migration based on access control logs, data preprocessing is performed on the access control decision log information, and after model training, the migration model containing historical access control log information is finally obtained; in the preparation step of the permission decision method for unstructured data based on semantic reasoning, all the permission texts are transformed into text vectors by using the language model in the under-chain indexing module. At the end of the preparation step, the subject user makes a cross-domain access control request for the object text resource, and the access request data contains the access request triad: subject, access action, and object. First, the smart contract parses the triad in the access request and uses the access control log migration model to make the first permission decision, and if the result is allow, the semantic reasoning method is used to make the second permission decision to get the final decision result. The smart contract is the access control information interaction hub, after authenticating the legitimacy of the access control request, two methods are executed successively to make a decision, and there is mutual collaboration between the on-chain and off-chain.

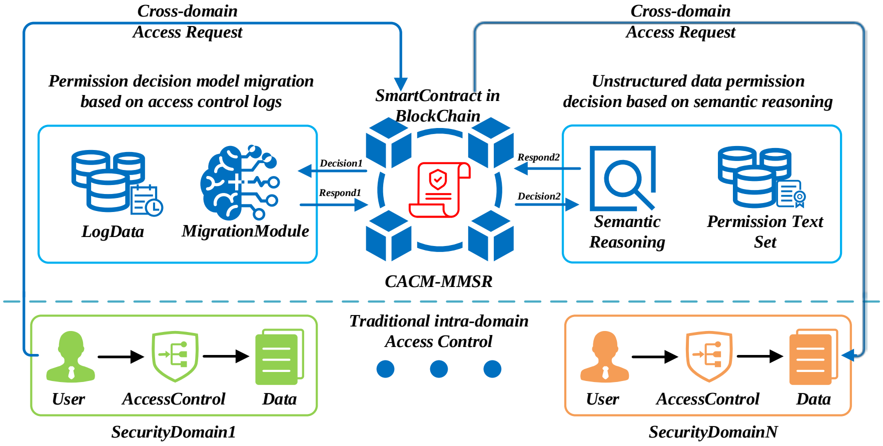

Cross-domain Access Request
Cross-domain Access Request
Permission decision model migration based on access control logs
SmartContract in BlockChain
Unstructured data permission decision based on semantic reasoning
LogData
MigrationModule
Decision1
Respond1
Respond2
Decision2
Semantic Reasoning
Permission Text Set
CACM-MMSR
Traditional intra-domain Access Control
User
AccessControl
Data
SecurityDomain1
User
AccessControl
Data
SecurityDomainN

Fig. 1. Overall procedure of CACM-MMSR mechanism.

4. Migration method for access control log-based permission decision model

4.1 Core Ideas

The access control log-based permission decision model migration method and detailed architecture are illustrated in Fig. 2. Each security domain undergoes permission decision model migration based on access control policy logs. The following steps are involved in different security domains:

1) Dataset Balancing and Target Coding. The dataset is balanced to enhance the performance of the training model. Additionally, target coding is applied to expedite the model training process.

2) ResNet Network Learning. The access control logs are learned using the ResNet network architecture. The model with superior performance metrics is selected as the final migration model.

3) Storage of Migration Model. The unique identification ID of the migration model is stored in the ID storage structure within the smart contract.

4) Migration Model-based Decision. The Migration Model-based Decision process algorithm is shown in Algorithm 1. When conducting a permission decision, the user's decision response is processed by the smart contract. The smart contract searches for the migration model of the access control object security domain in the ID storage structure. The migration model performs the decision and returns the decision result.

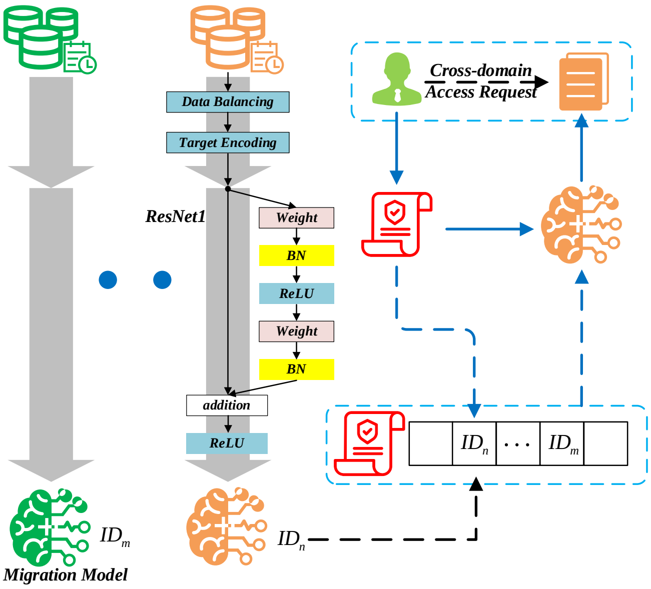


Fig. 2. Access Control Log-based Permission Decision Migration Model and Permission Decision Framework.

```
Algorithm 1: LogDecision
  Input: AccessControlRequest: sub,access,obj_hash
  Output: result
  Find MigrationModel with (sub,obj_hash) in ID Structure;
  Target encoding of request data;
  Get Respond from MigrationModel;
  if Respond is True then
      return True;
  else
      return False;
```

4.2 Access control log data processing

1)Data Balance

In systems with heterogeneous access control, access control logs often consist of structured data with a relatively fixed format, despite variations in access control mechanisms and policy languages. Leveraging neural networks to learn from extensive log data enables the modeling of relationships between different types of access control attributes and the patterns in decision results. However, real access control policy sets tend to be imbalanced, with significantly different proportions of access-allowing and access-denying policies. For example, in the access control policy set[10-11] published by Amazon.com, the ratio of access-allowing policies to access-denying policies is approximately 16:1. To address this, we adopt the Adaptive Synthetic Sampling Approach (ADASYN) to generate a balanced access control dataset, following these steps:

a) Use n_s to denote the number of policy samples allowed to be accessed and n_l to denote the number of policy samples prohibited from being accessed to calculate the unbalance degree:

$d = \frac{n_s}{n_l}$, $d \in (0,1]$.

b) Calculate the data $GN = \alpha \bullet (n_l - n_s)$, $\alpha \in [0,1]$ that requires the synthesis of samples. When $\alpha = 1$, GN is equal to the difference between the minority and majority classes, the data of the majority and minority classes in the synthesized dataset is exactly balanced.

c) Calculate the number of neighbors $r_i = \frac{N_l}{f}, r_i \in [0,1]$ for each minority f class sample using Euclidean distance, where N_l is the number of samples belonging to the majority class among the f neighbors.

d) Calculate the case of the majority class around the minority class sample $\tilde{r_i} = \frac{r_i}{\sum_{i=1}^{m_s} r_i}$, $\sum_{i=1}^{m_s} r_i = 1$.

e) Calculate the number of samples that need to be synthesized for each minority class sample $gn_i = \tilde{r_i} \cdot GN$.

f) Select one minority sample x_i from k neighbor around each minority sample x_{zi} to be synthesized for sample data synthesis: $s_i = x_i + \beta \cdot (x_{zi} - x_i)$, $\beta \in [0,1]$.

2)Data encoding

The target coding is to digitize the categorical variable through the target. This method replaces the categorical variable with a new numerical variable, and replaces each category of the categorical variable with its corresponding target probability. The mathematical expression for the target encoding is: $u = \frac{n \times \tilde{x} + m \times w}{n + m}$, where is the average value attempted to calculate, n is the number of access control logs, $\tilde{x}$ is the estimated average value, m is the weight assigned to the overall mean, and w is the overall average weight. m is the only parameter that needs to be selected. The higher the m, the more dependent it is on the overall mean of w . When $m = 0$, no smoothing is performed.

4.3 Model migration

We trained ResNet for model transfer by comparing various models. The convolution layer is used to convolution the input access control log data and provide a higher degree of feature representation. The convolution layer's parameters, known as the weights, are fundamentally tuned by backpropagation during training so that the network may discover the best feature representation. This procedure can be stated as follows:

$$F = W_{n+1}\sigma\left(W_n x\right) \quad (1)$$

Among them, skip connection in ResNet aids in maintaining information pertaining to input access control log data and transmitting it to higher network layers. By introducing skip connections, ResNet can train deeper network layers such as shallow neural network layers. The entire ResNet's output model can be modeled as:

$$MigrationModel = F\left(x, \left\{W_i\right\}\right) + x \quad (2)$$

5. A semantic reasoning based permission decision method for unstructured data

5.1 Core ideas

The detailed procedure of semantic reasoning based permission decision method for unstructured data is shown in **Fig. 3**. The overall flow algorithm of semantic reasoning based permission decision is shown in algorithm 2. After the user access control request is processed by the contract, IPFS_hash is first used to find the access control object resource and the user's permission text collection in IPFS. The permission text refers to the text resource defined for a user that already has access rights. The set of permission texts has been transformed into a set of permission text vectors by the language model in the offline indexing phase, and an indexed list structure has been created. Subsequently, the object text is semantically vectorized using the language model, and enters the semantic reasoning decision stage, which returns the final access control decision result after similar text retrieval and threshold decision.

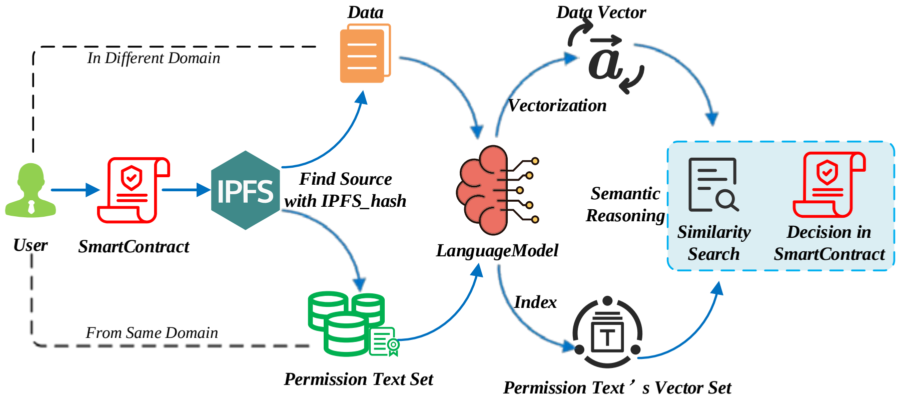


Fig. 3. Unstructured Data Permission decision Using Semantic Reasoning.

```
Algorithm 2: SemanticDecision
  Input: AccessControlRequest: sub,access,obj_hash
  Output: result
  Find obj from IPFS with obj_hash; // $obj\_hash \leftarrow obj$
  Vectorization and Index; // $obj \leftarrow Vector,\ text \leftarrow Vector$
  Function SemanticReasoning($Data\_Vector, Vector\ set$):
      Similarity Search;
      SemanticDecision;
      return $Respond$;
  Get Respond from SemanticReasoning;
  if $Respond$ is $True$ then
      return True;
  else
      return False;
```

5.2 Offline Indexing

1)Semantic vectorization

Language models can describe text at the semantic level and transform text into content-rich semantic text vectors. The language model MiniLM[12] used in this paper employs optimization methods such as Deep Self-Attention Distillation of knowledge to reduce the size and computational cost of the model. Deep Self-Attention Distillation has 3 core points. The first is to train the student model by deeply mimicking the "self-attention" module in the last layer of the teacher model, especially minimizing the KL difference between the teacher's and student's self-attention distributions:

$$\mathsf{L}_{AT} = \frac{1}{A_h \mid x \mid} \sum_{a=1}^{A_h} \sum_{t=1}^{|x|} D_{KL} \left(\mathbf{A}_{L,a,t}^{T} \cdot \mathbf{A}_{M,a,t}^{S} \right) \quad (3)$$

where $\mid x \mid$ and A_h denote the sequence length and the number of attentional heads, L and M denote the number of layers of teachers and students, and A_L^T and A_M^S are the attentional distributions of the last *Transformer* layer of teachers and students, respectively. The second point is that, in the self-attention module, the migration of relations between values (i.e., dot product operation between values) is introduced in addition to the conventional query-key dot product operation to achieve deeper imitation. The relation matrix *Values* is obtained by the dot product between *Values* vectors, which can represent the word-to-word dependency of *Values*. Taking the teacher-student value relationship $KL-divergence$ As a goal of training:

$$\mathbf{VR}_{L,a}^{T} = softmax\left(\frac{\mathbf{V}_{L,a}^{T} \mathbf{V}_{L,a}^{T\bullet}}{\sqrt{d_k}} \right) \quad (4)$$

$$\mathbf{VR}_{M,a}^{S} = softmax\left(\frac{\mathbf{V}_{M,a}^{S} \mathbf{V}_{M,a}^{S\bullet}}{\sqrt{d_k^{'}}} \right) \quad (5)$$

$$\mathsf{L}_{VR} = \frac{1}{A_h \mid x \mid} \sum_{a=1}^{A_h} \sum_{t=1}^{|x|} D_{KL} \left(\mathbf{VR}_{L,a,t}^{T} \parallel \mathbf{VR}_{M,a,t}^{S} \right) \quad (6)$$

Among them, $\mathbf{V}_{L,a}^{T} \in R^{|x| \times d_k}$ and $\mathbf{V}_{M,a}^{S} \in R^{|x| \times d_k^{'}}$ are the attention header values for the last layer of attention modules for teachers and students. $\mathbf{VR}_{L}^{T} \in R^{A_h \times |x| \times |x|}$ and $\mathbf{VR}_{M}^{S} \in R^{A_h \times |x| \times |x|}$ are the value relationships of the last layer *Transformer* for teachers and students, respectively. The third argument is that the addition of a teaching assistant aids in model distillation when there is a significant size gap between the teacher model and the student model. Compress the enormous model first into a medium-sized model, then use this medium-sized model as a "teaching assistant" to further compress it into the ultimate compact model.

2)Index establishment

The closer two text vectors are after being vector transformed with a language model, the more similar the texts are, i.e. the formula that the highest similarity value satisfies is $Similarity(x) = \arg\min_{y \in Y} d(x, y)$. We used the IVFPQ (Inverted File with Product Quantization)[13] indexing technique to swiftly locate the object text that is most similar among large permission texts. Finding the sum of squares of the differences between the object text vector and the authority text vector after vector quantization is how the IVFPQ index calculates text similarity. Formula expressed as follows:

$$\tilde{d}(x, y) = d(x, q(y)) = \sqrt{\sum_{j} d(u_j(x), q_j(u_j(y)))^2} \quad (7)$$

In the process of index establishment, the steps for permission text set Y, permission text $y \in Y$, and indexing vector y are:

a) Using Y clustering for $K-means$, obtain the rough quantified class center q_c, and record the number of samples for each class and the category to which each sample belongs. The number of centers in this class is the number of converted lists. Save all class centers in a table called coarse_cluster table, where each item is d dimensional.

b) Calculate the margin $r(y) = y - q_c(y)$ of y. $r(y)$ dimension is the same as y ,and then all $r(y)$ features are divided into m group using product quantization. Within each group, $K-means$ clustering is still used, and the result is a m dimension vector. Save all product quantization results in a table called pq_centroids table, each item in the table is m dimension, $y \approx q_c(y) + q_p(r(y))$.

c) Record index i of y in coarse_cluster table and index j in pq_centroids table. When inserting the inverted list, insert id, j) into the i th inverted index L_i. id is the identifier of y. The length of the list is the number of sample y belonging to class i.

5.3 Semantic Reasoning

1) Similar text retrieval

After establishing the permission text index, the retrieval steps for object text x in permission text set Y are:

a) Roughly quantize x, that is, use the KNN method to classify x into a certain class or several classes.

b) Calculate the remainder of x as $r(x)$, which can be expressed as: $x = q_c(x) + r(x)$.

c) Maximum Heapsort. Each element in the heap represents the distance $\| x - y \| = \| q_c(x) + r(x) - q_c(y) - q_p(r(y)) \| = \| r(x) - q_p(r(y)) \|$ between y and x in the permission text set. The distance of the top element of the heap is the largest. As long as the element is smaller than the top element of the heap, replace the top element of the heap and adjust the heap until all y 's are decided. Finally, the nearest distance, that is, the highest similarity value $Similarity = \min_{\{y|q_c(y)=q_c(x)\}} \| x - y \|$, is obtained.

2) Threshold decision

A similarity score between 0 and 5 was used to reflect the resemblance relationship between two assertions in their study of this topic. Between them, 0 denotes that there is no semantic relationship and 5 denotes that their meanings are identical. Based on the statement's similarity score level and the scoring criteria, we divide the threshold for similarity assessment. **Table 1** provides language examples at five threshold levels along with the example. Based on the statement's similarity decision threshold, we can manually set a fine-grained threshold method that satisfies the requirements.Our decision threshold can also be set between 0 and 5 on smart contracts in order to provide a more precise choice procedure, and the decision formula is:

$$Respond(request) = \begin{cases} 1, Similarity \geq threshold \\ 0, Similarity < threshold \end{cases} \quad (8)$$

The blockchain's consensus method is used to carry out the threshold decision, assuring the validity of the decision.

6. Experiment and analysis

To prove the uniqueness and effectiveness of CACM-MMSR, we designed three experiments. Firstly, section 6.1 proves the efficiency improvement of blockchain-based distributed access control system by designing different number of nodes to handle the access requests. Second, section 6.2 demonstrates the accuracy of the access control based on model migration approach by analyzing the performance of the migration model on amazon-employee-access-challenge data. Finally, section 6.3 demonstrates the accuracy and high efficiency of access control based on semantic reasoning by comparing the performance of different vectorization and indexing methods. The data set used in the experiment is shown in Table 1.

Table 1. The data set used in the experiment

	Name	Experiments	Description
1	Distributed access requests dataset	6.1	We constructed 1000 access control requests containing 10 user types, 3 user actions, and 50 different object resources to invoke the smart contract for decision making.
2	Amazon employee access control policy dataset	6.2	The data consists of real historical data collected from Amazon. This dataset includes more than 32,000 real access control policy messages.
3	Semantic textual similarity benchmark dataset	6.3	The STS-B dataset contains 5,749 pairs of sentences, each of which has a subjective human score indicating the similarity between the two sentences, ranging from 0 to 5, where 0 indicates two sentences that are not related and 5 indicates a sentence that is exactly the same.

6.1 Distributed access control decision performance analysis

We simulate multiple peers invoking the smart contract decision function in the Fabric blockchain environment and calculate the time when all access control requests are processed, thus simulating the time it takes for different users to send access control requests in a distributed manner until the requests are processed, to test the usability of the mechanism under distributed concurrent requests. The experimental results are analyzed to show that with a certain total size of the decision requests input to CACM-MMSR, the number of nodes increases, and the time to finish processing all the decision requests is gradually shortened and eventually stabilized. The experimental results show that the use of blockchain smart contracts for distributed processing of concurrent access control requests can effectively improve the efficiency of access control decisions in distributed environments, and the final time of decision requests tends to stabilize because it takes a certain amount of time to carry out consensus among blockchain nodes. The result shows that distributed access control decision time expense of CACM-MMSR can be reduced to less than 28.7% of a single peer. The results are shown in Fig. 4.

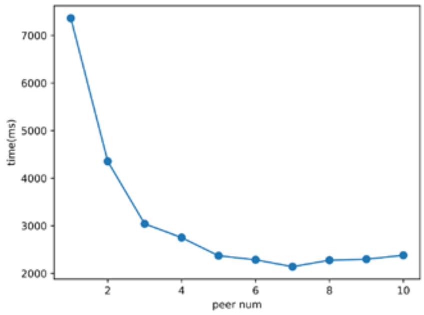


Fig. 4. For a Certain Size of Total Decision Requests (1000) The Relationship Between the Number of Different Peers and the Time Taken to Finish the Decision.

6.2 Analysis of permission decision model migration performance

In order to verify the performance and efficiency of the access control log-based permission decision model migration method proposed in this paper in a novel computing environment, we conduct experiments using the real access control policy set[10-11] from Amazon.com as the decision log information, which is a commonly used dataset for evaluating access control. This dataset includes more than 32,000 real access control policy messages, including 10 different categories of user attribute information, as shown in **Table 2**. We performed data balancing on the experimental dataset and randomly partitioned the data with a ratio of 4:1 between the training dataset and the test dataset.

Table 2. Description of the Access Control Log Data Set

Attribute items	Description	Number of categories
ACTION	Permission marker (allow or deny)	2
RESOURCE	Resource ID	7518
MGR_ID	The specific manager ID for the present employee	4243
ROLE_ROLLUP_1	the first type of roles	128
ROLE_ROLLUP_2	the second type of roles	177
ROLE_DEPTNAME	Department Role Categories	449
ROLE_TITLE	Employment Type	343
ROLE_FAMILY_DESC	Extension of the role family description	2358
ROLE_FAMILY	description of role family	67
ROLE_CODE	Role code that exclusively matches the role category	343

We used this dataset to train several different neural network models as decision models, and used a test set to evaluate the decision accuracy of several models. When evaluating the access control decision accuracy of the migration model, we define the confusion matrix of permission decision results as follows in **Table 3**.

Table 3. Results of the confusion matrix for the privilege decision

Reality	Predicted results	
	Access allowed	Access denied
Access allowed	$D_{PP'}$	$D_{PD'}$
Access denied	$D_{DP'}$	$D_{DD'}$

Where TP represents the number of samples that were correctly determined to be granted access, FN represents the number of samples that were incorrectly determined to be denied access, FP represents the number of samples that were determined to be granted access but were denied access, and TN represents the number of samples that were determined to be granted access but were actually granted access. The corresponding assessment criteria are Accuracy $Acc = (D_{PP'} + D_{DD'}) / (D_{PP'} + D_{PD'} + D_{DP'} + D_{DD'})$, Precision: $Pre = D_{PP'} / (D_{PP'} + D_{DP'})$, Recall: $Re = D_{PP'} / (D_{PP'} + D_{PD'})$, F1 Score: $F1 = (2 \cdot Pre \cdot Re) / (Pre + Re)$.

1) Accuracy Analysis

In order to evaluate the performance of the methods, Accuracy, Precision, Recall and F1 Score of KNN, SVM, Random Forest, MLP, CNN, ResNet are compared under the condition of selecting the same features. The results are shown in **Fig. 5**.

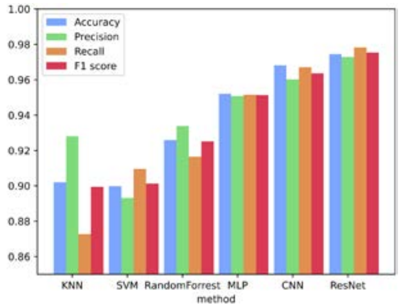
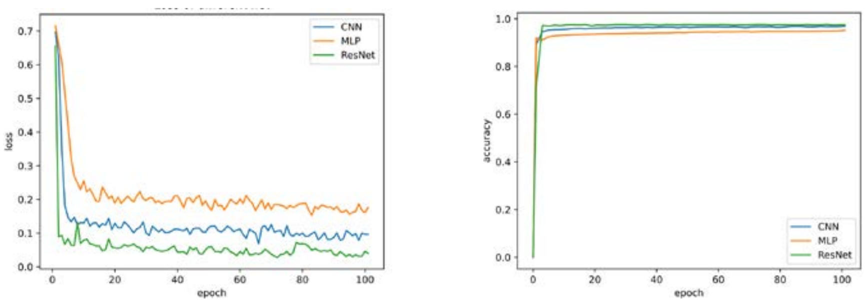


Fig. 5. Comparison of the Results of Several Techniques.

The experimental results show that the neural network class methods are more accurate than other machine learning algorithms, which has a high decision accuracy of 97.4%. The loss function values and accuracies of the three neural network trainings as a function of the number of trainings are shown in **Fig. 6**. In summary, the ResNet used in this paper has a better performance for comprehensive permission decision.



(a)Train Loss (b)Train Accuracy

Fig. 6. Relationship Between the Number of Training Sessions and the Accuracy of the Three Neural Networks on the Test Set.

2) Analysis of time expense

In order to evaluate the time overhead of the method, three neural networks are used for permission decision modeling for access control log migration.The ResNet model has essentially the same decision time with improved accuracy relative to the other two network models. As the resource size grows, the permission decision time grows positively correlated, and access control decisions can be made in less than 5 seconds for a resource size of 100,000, and in milliseconds when the resource size is in the 10,000 range or less. The method in this paper requires shorter decision time, has better overall performance, and can better adapt to the access control needs of real-time decision of massive policies.

6.3 Performance evaluation of authorization decisions for unstructured data

Since the core of the semantic reasoning-based permission decision method lies in the semantic vector transformation of text and semantic reasoning, we design experiments that use several different language models for vectorization, as well as several different indexing methods to index these vectors, and test the performance of these different methods. The results are shown in **Fig. 7**.

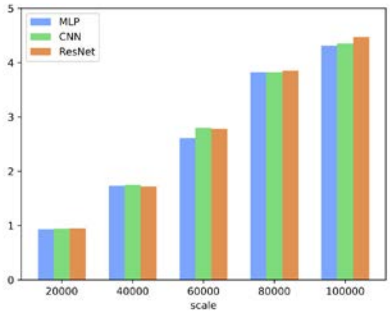


Fig. 7. Decision-making Costs for Various Models and Resource Sizes.

1) Performance evaluation of semantic vectorization

In order to verify the performance of the mechanism in semantic vectorization of object resources, we conducted experiments using the STS dataset [14] with Pearson's correlation coefficient and Spearman's correlation coefficient as the performance metrics to test the ability of the different semantic models to express the semantics. The formulae of Pearson's correlation coefficient are as follows:

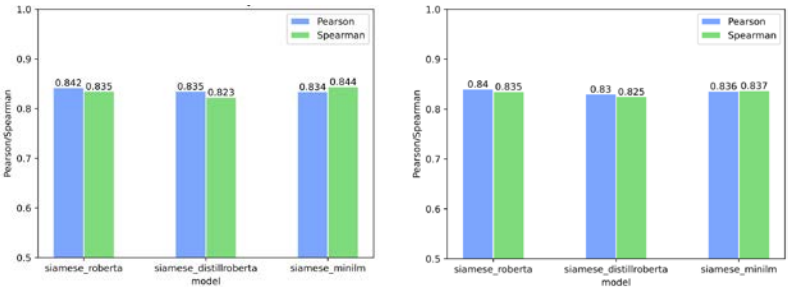
$$r_{xy} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2}\sqrt{\sum (y_i - \overline{y})^2}} \quad (9)$$

r_{xy} denotes the Pearson correlation coefficient, x_i and y_i represent the relative values of the two variables from the i sample. $\overline{x}$ and $\overline{y}$ represent the corresponding means of the two variables. The linear relationship between two variables is described by r_{xy}, which accepts values between $[-1,1]$. When r_{xy} is closer to 1, it denotes a greater and more positively correlated linear link between the two variables. There is no linear association between the two variables when $r_{xy} = 0$.

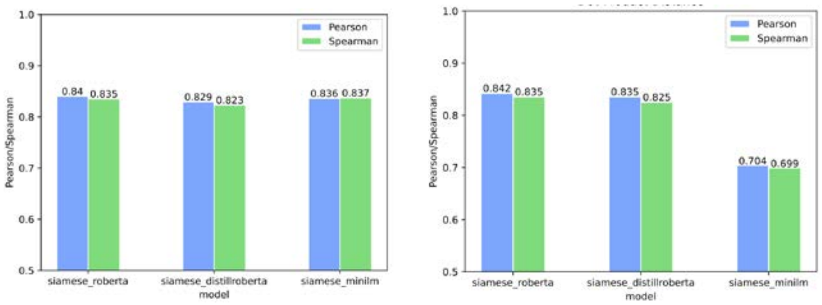
The Spearman's rank correlation coefficient (SRC) has the following formula:

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \qquad (10)$$

r_s represents the Spearman correlation coefficient. n represents the sample size, d_i represents the ith sample's rank difference. The monotonic relationship between two variables is described by Spearman's correlation coefficient, a non-parametric correlation coefficient. The results are shown in **Fig. 8**.



(a) Cosine-Distance (b) Euclidean-Distance



(c) Manhattan-Distance (d) Dot-Product-Distance

Fig. 8. Performance of Twin Network Topologies Using Various Distance Calculation Techniques for Three Language Models.

When the same training set, test set, and validation set were chosen, we calculated the similarity, validation results, and Pearson correlation coefficients of the test set using various similarity decision methods to test the accuracy of various language models. For testing, we created a siamese network architecture based on the language models Roberta[15], DistilRoberta[16], and MiniLM. For the three language models and various authority text sizes, we evaluated the time overhead associated with semantic vectorization. The efficiency of Roberta is worse than DistilRoberta and MiniLM, and the gap is obvious. On the other hand, DistilRoberta's efficiency is not as good as MiniLM's, but the gap is not as obvious. The semantic vectorization time of the three language models is positively correlated with the text size. Therefore, in terms of efficiency, DistilRoberta and MiniLM are more suitable for this paper's research. The results are shown in **Fig. 9** and **Table 4**.

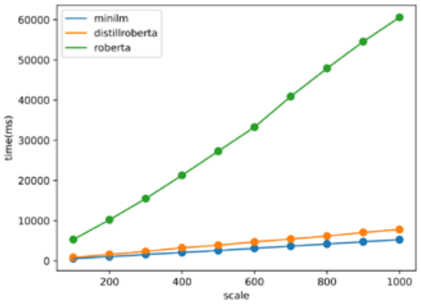


Fig. 9. The Amount of Time Spent Semantically Vectorizing the Three Language Models with Various Permitted Text Sizes.

Table 4. Number of parameters for the three language models

Model	Param. (Millions)
all-roberta-large-v1[17]	360
all-distilroberta-v1[18]	67
all-MiniLM-L6-v2[19]	0.285

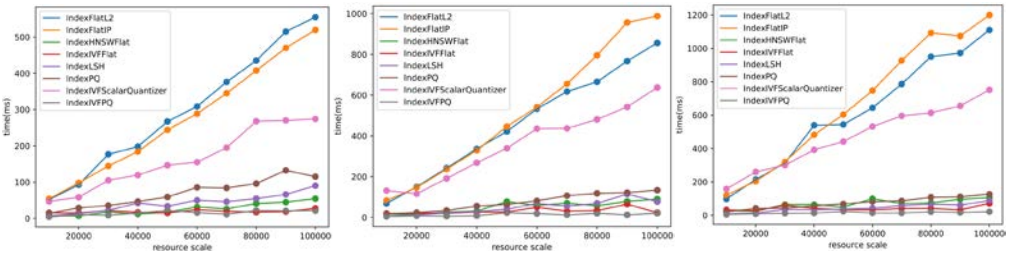
In semantic reasoning-based permission decision methods for unstructured data, the choice of language model determines the method's ability to semantically express the textual resources, which in turn affects the performance of semantic reasoning-based decision methods. In a big data environment, we often think of using large-scale pre-trained models such as Bert[20] or RoBerta. In the process of comprehensive comparison, the number of model parameters of MiniLM is 1/235 of that of BERT, the inference speed is relatively fast, especially compared with DistilRoberta has a substantial improvement, and in the process of model accuracy test, the similarity is calculated by using the cosine distance, the Euclidean distance, and the Manhattan distance, and it still maintains the similarity with Roberta and DistilRobertad with almost the same accuracy, therefore, in summary, using MiniLM as the language model for this method is a relatively superior choice.

2) Analysis of Similar Text Retrieval Performance

The accuracy of semantic reasoning is mainly related to the threshold set in the threshold decision stage, which is subjective, so the performance analysis of semantic reasoning mainly focuses on the time overhead. In semantic reasoning, the major influence on time overhead is the similar text retrieval stage, so this section mainly focuses on similar text retrieval for the analysis of time overhead. The time required for similar text retrieval of object text vectors in the index table is mainly determined by the way the index is built. Secondly, the time taken by text vectors of different dimensions to perform similar retrieval varies, and text vectors of higher dimensions take more time to retrieve on the basis of different indexing ways. In order to find the optimal way of index building, we use Faiss[21] for index building, the specific method and introduction are shown in **Table 5**.

Table 5. Methodology and search creation introduction

Method	Introduce
IndexFlatL2	Exact Search for L2
IndexFlatIP	Exact Search for Inner Product
IndexHNSWFlat	Hierarchical Navigable Small World graph exploration
IndexIVFFlat	Inverted file with exact post-verification
IndexLSH	Locality-Sensitive Hashing (binary flat index)
IndexPQ	Product quantizer (PQ) in flat mode
IndexIVFScalarQuantizer	IVF and Scalar quantizer (SQ) in flat mode
IndexIVFPQ	IVFADC (coarse quantizer+PQ on residuals)



(a) 384 Dimensions (b) 768 Dimensions (c) 1024 Dimensions

Fig. 10. Time Costs Associated with Various Indexing Techniques for Resources of Various Sizes and Text Vector Dimensions.

The results are shown in **Fig. 10**. The experimental results show that the use of IVFPQ index building method for similar text retrieval, both in the change of data size and dimensional changes, the time overhead is relatively small and insignificant changes, is the more ideal index building method in this paper. The semantic reasoning-based permission decision method maintains similar performance to the comparison model in three common methods for computing vector similarity such as Cosine-Distance, while using MiniLM is 6.4% of that of roberta and 57% of that of distilroberta in the vectorization time.

7. Conclusion

In this paper, a cross-domain access control mechanism is designed by combining blockchain and deep learning, which provides ideas for solving cross-domain access control in new computing environment. The access control decision function is implemented on smart contracts, which is well in line with the cross-domain environment in the open environment. The access control log-based permission decision model migration method and the semantic reasoning-based permission decision method for unstructured data are implemented based on neural networks, which can effectively analyze the massive data resources in a fine-grained and automated way in the big data environment. Moreover, the permission decision model migration method based on access control logs can not only solve the problem of access control mechanism heterogeneity among different domains, but also solve the problem of access control policy compatibility between old and new access control mechanisms in the same security domain. Finally, we verify the effectiveness of the proposed method through experiments. Since the object of access control in this paper is mainly text data, we will study the access control of image and video data in the future work.

Acknowledgement

This work is supported by Key Research and Development and Promotion Program of Henan Province (No.222102210069), Zhongyuan Science and Technology Innovation Leading Talent Project (224200510003), and National Natural Science Foundation of China (No. 62102449).

References

[1] L. Fang, L. H. Yin, Y. C. Guo, and B. C. Fang, "A Survey of Key Technologies in Attribute-Based Access Control Scheme," *Chinese Journal of Computers*, vol. 40, no. 07, pp. 1680-1698, Jul. 2017.

[2] W. C. Wu, Z. Y. Ren and X. H. Du, "Log-based rich-semantic ABAC policy mining," *Journal of Zhejiang University (Engineering* Science*)*, vol.54, no.11, pp.2149-2157, 2020. Article (CrossRef Link).

[3] A. Liu, X. H. Du, N. Wang, and R. Qiao, "ABAC Access Control Policy Generation Technique Based on Deep Learning," *Journal of Communications*, vol.41, no.12, pp.8-20, Dec. 2020. Article (CrossRef Link).

[4] J. B. Zhang, Z. Q. Zhang, W. S. Xu, and N. Wu, "A Blockchain Based Interdomain Access Control Model," *Journal of Software*, vol.32, no.5, pp.1547-1564, 2021. Article (CrossRef Link).

[5] H. Arshad, C. Johansen, and O. Owe, "Semantic Attribute-Based Access Control: A review on current status and future perspectives," *Journal of Systems Architecture*, vol.129, 102625, Article (CrossRef Link).

[6] A. Jawed Khan and S. Mehfuz, "Fuzzy User Access Trust Model for Cloud Access Control," *Computer Systems Science and Engineering*, vol.44, no.1, pp.113-128, 2023. Article (CrossRef Link).

[7] F. Sabrina and J. Jang-Jaccard, "Entitlement-Based Access Control for Smart Cities Using Blockchain," *Sensors*, vol.21, no.16, 5264, 2021. Article (CrossRef Link).

[8] K. Ma and G. Yang, "RTBAC: a Risk-Aware Topic-Based Access Control Model for Text Data with Paragraph-Level Authorization," *Security and Communication Networks*, vol.2022, no.1, 2022. Article (CrossRef Link).

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.770-778, 2016. Article (CrossRef Link).

[10] Access Control Dataset. [Online]. Available : https://www.kaggle.com/c/amazon-employee-access-challenge/data.

[11] L. Karimi, M. Aldairi, J. Joshi, and A. Mai, "An Automatic Attribute-Based Access Control Policy Extraction From Access Logs," *IEEE Transactions on Dependable and Secure Computing*, vol.19, no.4, pp.2304-2317, 2022. Article (CrossRef Link).

[12] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers," in *Proc. of MINILM: deep self-attention distillation for task-agnostic compression of pre-trained transformers*, vol.8, p p.5776-5788, 2020. Article (CrossRef Link).

[13] H. Jégou, M. Douze, and C. Schmid, "Product Quantization for Nearest Neighbor Search," *IEEE Transactions on Pattern* Analysis *and Machine Intelligence*, vol.33, no.1, pp.117-128, 2011. Article (CrossRef Link).

[14] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation," in *Proc. of 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp.1-14, 2017. Article (CrossRef Link).

[15] Y. Liu, M. Ott, N. Goyal, J. F. Du, M Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv.org*, 2019. Article (CrossRef Link).

[16] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv.org*, 2020. Article (CrossRef Link).

[17] all-roberta-large-v1. [Online]. Available: https://huggingface.co/sentence-transformers/all-roberta-large-v1

[18] all-distilroberta-v1. [Online]. Available: https://huggingface.co/sentence-transformers/all-distilroberta-v1

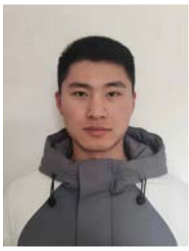
[19] all-MiniLM-L6-v2. [Online]. Available: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv.org, 2018. Article (CrossRef Link).

[21] J. Johnson, M. Douze, and H. Jegou, "Billion-scale similarity search with GPUs," *IEEE Transactions on Big Data*, vol.7, no.3, pp.535-547, Jul. 2021. Article (CrossRef Link).

TAN Ming, born in 2001, Ph. D. candidate. His current research interests focus on AI security and data security.

LIU Ao-Di, born in 1992, Ph. D., lecturer. His current research interests focus on blockchain security.



WANG Xiao-Han, born in 2000, Ph. D. candidate. His current research interests focus on big data security.

SHANG Si-Yuan, born in 2000, Ph. D. candidate. His current research interests focus on blockchain security sharing.

WANG Na, born in 1980, Ph. D., associate professor. Her current research interests focus on network and information security.

DU Xue-Hui, born in 1968, Ph. D., professor. Her current research interests focus on network and information security.