IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 5, May 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105243 |

# Research on Intelligent Security Protection of Privacy Data in Government Cyberspace

**Sajid. M. Momin[1], Dhanaji V. Mirajkar[2], Namrata N.Patil[3], Chaitanya Kulkarni[4]**

Faculty, Rajarambapu Institute of Technology, Islampur, Sangli, India[1]

Faculty, Rajarambapu Institute of Technology, Islampur, Sangli, India[2]

Faculty, Rajarambapu Institute of Technology, Islampur, Sangli, India[3]

Faculty, Rajarambapu Institute of Technology, Islampur, Sangli, India[4]

**ABSTRACT**- Based on the analysis of the difficulties and pain points of privacy protection in the opening and sharing of government data, this paper proposes a new method for intelligent discovery and protection of structured and unstructured privacy data. Based on the improvement of the existing government data masking process, this method introduces the technologies of NLP and machine learning, studiestheintelligentdiscoveryofsensitivedata,theautomatic recommendation of masking algorithm and the full automatic executionfollowingtheimprovedmaskingprocess.Inaddition, the dynamic masking and static masking prototype with text and database as data source are designed and implemented with agent-based intelligent masking middleware. The results show that the recognition range and protection efficiency of government privacy data, especially governmentunstructured text have been significantlyimproved.

**KEYWORDS**-categorization and classification; sensitive data discovery, data masking algorithm, NLP, Machine learning.

## I.INTRODUCTION

At present, there is no specific law to define sensitive data and user privacy in China, and the norms and standards of categorization and classification of government data are being established, which can't effectively identify important data, sensitive data and privacy data, which seriously affects the opening and sharing of government data. In daily work, there are many specific difficulties and pain points in data privacyprotection:The Requirements of data masking are not clear. When applying for data masking services, most government departments are difficult to accurately describe the needs, including which fields to keep, which attributes to keep, which fields must be masked, which fields' statistical information must be kept or hidden (such as distribution, mean change, constant total or classification attributes, and other information containing correlation), and the masking algorithm used in specific fields. Usually the requirements are simple, general andunclear.

*1)* Sensitive information is not easy to define. When data providers sensitive data for sharing, they will focus on data security and compliance, hoping to provide data in accordance with the principle of minimization; however, data users want to get more and more complete data, which urgently needs to refine the definition of sensitive data. For example, generally the personal information ofcitizens related, data masking should be carried out in principle, but sensitive data can't be simply understood as basic personal information. Because of the amount of data, information in association tables, and even seemingly non critical information related to other data, it is possible to form or infer global or individual information, resulting in sensitive information leakage.

2)The balance between data protection and data availability. Data masking inevitably leads to data information loss. On the one hand, choosing appropriate masking rules or methods requires better masking software and proficient mastery of various tools provided by the software; on the other hand, the data demand department needs to fully participate in and fully understand the masking work, and jointly find the balance between the protection and availability of sensitivedata.

3)The efficiency of manual configuration, rule-based sensitive data discovery and masking algorithm recommendation is getting lower and lower. With the coming of big data era, the amount of data is increasing rapidly and new sensitive data types are emerging in an endless stream, as well as the urgent application of a large number of government unstructured data, which makes the manual configuration and rule-based identification more

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105243 |

and more time-consuming and difficult. As a result, a large number of sensitive data cannot be processed, which seriously affects the efficiency of desensitization and the effect of privacyprotection.

This paper proposes a method of accurate location and intelligent security protection of privacy data in government cyberspace. On the basis of automatic categorization and classification of government data, sensitive data is automatically discovered from big data according to specific business scenarios, and then desensitization strategies and algorithms are automatically recommended in combination with sensitive data types, so as to improve the overall recognition rate and de sensitization efficiency of sensitive data.

## II. WORK-FLOW IMPROVEMENT

A full work-flow of data masking includes finding sensitive data, determining masking methods, defining masking rules, performing masking operations and evaluating masking effect. This research introduces unstructured data, andaccording to the characteristics of specific business scenarios and sensitive data, uses artificial intelligence, machine learning and other technologies to reform the existing process. So we get the following improved data maskingwork-flow.

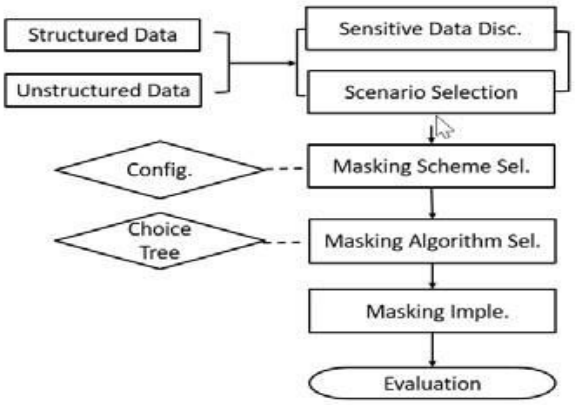


Figure 1. Improved Data Masking Work-flow

1.InputData -The proportion of the unstructured data is more than 80%, including text, audio, picture and video, which contains alarge in number of sensitive information, a wide variety anddifficult to process. This paper focuses on the recognitionand protection of sensitive data in text.

2.Automatic Discovery of SensitiveData

In order to implement differentiated data privacyprotection on by sin artificial intelligence technology the dataaccurately class the data is accurately classified and the data to be desensitized and its attributes are determined in combination with specific business scenarios and data characteristics; sensitive data is automatically discovered and the sensitive level of sensitive data is determined through rule matching, NLP and artificial intelligence modelling. The auto-discovery prepares for different masking schemes and the recommendation of diversified masking algorithms.

3.Diversified masking algorithms-

Reasonable categorization of data resources is an important technical means to improve the efficiency and data availability, by using the method of multi-dimensional and l ear classification, the government data are classified in three dimensions: theme, industry and service. For each dimension, the linear classification is used to classify it into three levels: large category, medium category and small category. Business departments can classify data into sub categories according to business needs. For the subdivision of small categories, each department can expand and subdivide them according to the nature, function and technical means of business data.

In this paper, two new categories, domain and sector are added. Domain categorization is the thematic categorization and packaging of all data according to the actual situation of government data and local characteristics. Sector categorization is the categorization based on the administrative function settings of sectors as shown in the following figure.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105243 |

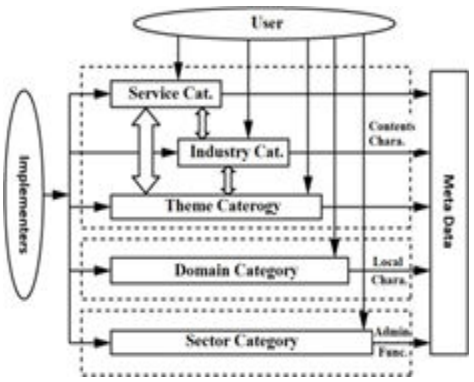


Figure 2. Government Data Categorization

Directivity refers to the range of data that can be related to specific objects, which can be divided into a single subject (individual or institute), a specific group (investor category, listed company industry) or the whole market. Importance, privacy and directivity are expressed in three levels: high, medium and low. The data sensitivity level can be divided into four levels, which are determined comprehensively according to the importance, privacy and directivity of data, and fully consider the influenceof data quantity (full amount/ sampling), data association and data timeliness on data sensitivity.

II.Business ScenarioSelection

Considering the network environment, the business application scenarios of data masking can be divided into fourcategories

*1)* Internal Analysis. In the same kind of business network, the data analysis is carried out by analysts using masking data.

*2)* System Simulation. The network where the simulation experiment is carried out belongs to the same network environment as the system running environment, andmost of the users are developers and testers.

*3)* Regulatory Collaboration. The masking data will be provided to other regulatory cooperation agencies for use in their business networks, and the users are regulatory business personnel.

*4)* External Analysis. It is used by analysts in the internet environment. Compared with the internal analysis, the network environment is open and the security level is low.

III.Masking SchemeSelection

For the application scenarios of sensitive data, the masking schemes can beselected following

1)Static Data Masking: after masking the original data once, the result data can be used many times, which is very suitable for the situation of single usescenario.

2)Dynamic Data Masking: it is a data masking for processing display data according to different user requirements when displaying sensitive data. It requires the system to have security measures to ensure that users can not directly contact sensitive data by bypassing the data masking. Dynamic data masking is more suitable for the situationofuncertainuserneedsandcomplexusescenarios.

A)Masking AlgorithmSelection

The biggest difficulty of data masking is to balance the privacy protection and data availability. Whether the masking algorithm is appropriate or not directly affects the desensitization effect. In order to develop an appropriate masking algorithm, the following factors are mainly considered in combination with specific application scenarios:

1) Availability. The desensitized data should meet the needs of analysis and application. If the desensitized data cannot be used for target analysis and application, it has no use value. In a specific application scenario, it may be necessary to retain some non-key information (such as ID number, some fields of mobile phone number, etc.) to meet the analysisrequirements.

2) Relevance. In the same data table, a field corresponds to another field. If masking algorithm breaks this relationship, the value of this field will no longer exist. Generally, when the reference is needed for data

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105243 |

statistics, the relevance of data is required to behigh.

3) Authenticity. The degree of preservation of the desensitized data to the logical characteristics and statistical distribution characteristics of the original data. To meet this feature, the original value of the data needs to be preserved as much aspossible.

4) Timeliness. Data provision needs to be timely. After a certain period of time, desensitization data may no longer have the significance of further analysis and mining. Therefore, we should try to avoid usingtime-consuming Desensitization algorithm, such as encryption algorithm.

• Reproducible. When the same source data is configured with the same algorithm and parameters, the desensitized data should be consistent, and random algorithm should beavoided.

• Configurable. It can flexibly configure and organize masking algorithms, and generate personalized desensitization data according to different requirements.

## III. DESIGN ANDIMPLEMENTATION

This chapter introduces the design and implementation of the system in detail following the improved work-flow above.

1. System architecture

Artificial intelligence model, it realizes the functions of sensitive data identification, data masking, and scheduling and monitoring of masking tasks. In addition, the management functions include: data source configuration, masking scene selection, dynamic and static masking selection, masking taskmanagement.

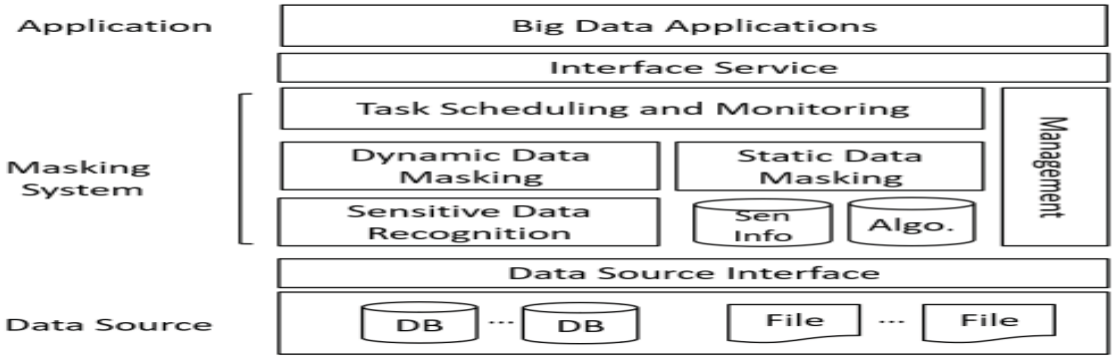


Figure 3. System Architecture

The upper layer big data application calls the URL interface to operate the intelligent data masking middleware and the operation of data acquisition. The main calling process is asfollows:

First, by calling the URL interface through zookeeper registry center, the user obtains the interface service address and enters a single service interface. Secondly, through the role user resource management to authenticate the user connection information, and to judge the authority. Then, operate the data masking middleware according to the user authentication authority information, interface function and parameters; If a data acquisition process, it will connect the data source through the data source interface information, and obtain the required data from it, start the intelligent sensitive data identification model, and isolate the sensitive data and attributes.

Finally, according to the characteristics of the data, start the intelligent masking algorithm recommendation model to select the most suitable algorithm then implement sensitive datamasking.

*A.* DataSource

The source data is divided into structured data and unstructured data. The latter includes text data, pictures, voice and video data. Generally, structured data exists in the database and unstructured data exists in the file (txt, html, pdf, Microsoft office format,etc.).

*B.* Library

*1)* Sensitive Feature Lib: after obtaining the feature data of text, audio and image video through training set, the security department and business personnel will identify and classify the corpus and feature database together, and select the representative words, image blocks and audioframes that can be identified by as sensitive information.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105243 |

| Name | DeSCription | Name | DeSCription |
|---|---|---|---|
| Encryption | Convert to meaningless value | Enumeration | Map to new values keeping data order |
| Hiding | Replace with a constant; this field is not required | Truncation | Truncate data tail |
| Hashing | Map to hash value; set indefinite length data to fixed | Prefix-preserving | Keep the first n bits of IP and confuse the rest |
| Permutation | Map to unique value | Mask | Keep the size, keep part of the info. |
| Shift | Add a fixed offset; hide some features | Floor | To round date/number |

Table 1 Masking rules oralgorithms

V.Agent-based Intelligent MaskingMiddleware-

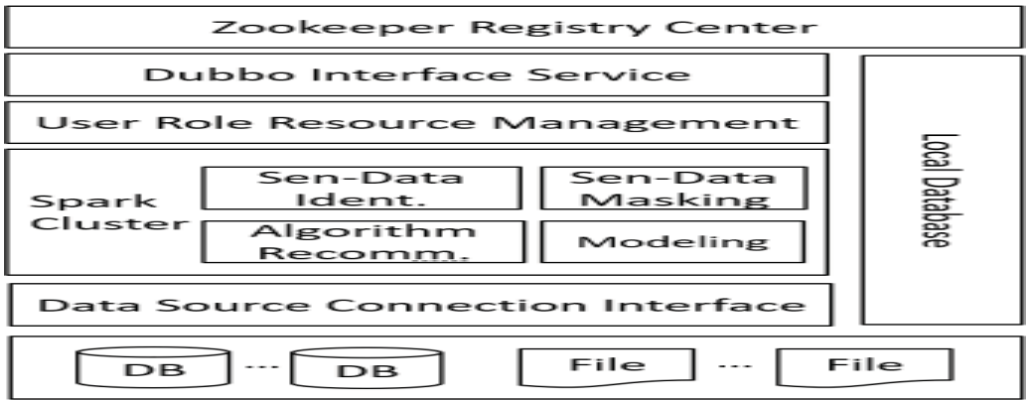


Figure 4. Intelligent Masking Middleware

The architecture of intelligent data masking middleware shown in the figure above, and the specific technical ascription is as follows.

*1)* Dubbo and Zookeeper Framework Integration: in order to meet the needs of high scalability and efficiency, RPC remote call service is used to provide user interface. The framework of Dubbo and Zookeeper is integrated to meet the requirements of high-performance and transparent interface process call, and the multi-cluster distributed interface call task is realized to achieve software load balancing, and the interface URL service management is optimized. After framework integration, middleware supports multiple protocols, multiple thread models, and directory services in the form ofregistry.

*2)* User Authentication and Data Permission Control: configure with Shiro framework to realize user authentication and multiple user data permission settings in interface service. Shiro can manage and configure users, roles and data resources, and different user role groups can have different data permissions and give data permissions to users.

*3)* Data Processing: use Spark parallel computing framework, spark SQL for interactive query, and Java RDD for data processing. Through Spark cluster, machine learning modeling, sensitive data discovery,masking algorithm recommendation and data batch masking are carried out for massive dataefficiently.

The above middleware acts as an agent, intercepts theirs data request in real time, and desensitizes therequested and return

Text Sensitive DataDiscovery :-

The process of text sensitive data recognition includes text categorization, text preprocessing and sensitive word recognition and sensitivity level classification.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105243 |

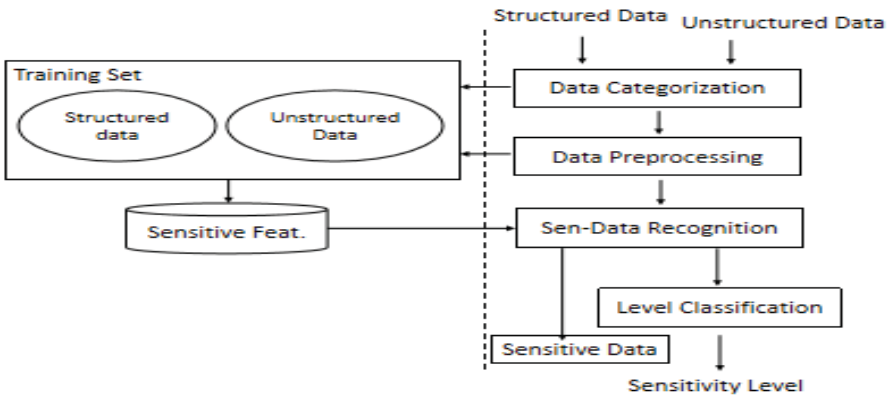


Figure 5. Intelligent Discovery of Sensitive Data

Text Categorization: through docking with the government data categorization subsystem, we can get the categorized text. These categories include large, medium, small, and more detailedsub-classes.

*2)* Text Preprocessing: Based on string matching algorithm combined with segmentation dictionary and word sequence, optimize the forward maximum matching algorithm to achieve accurate segmentation. The results usually have quality problems and need to becleaned.

*3)* Recognition and Grading of Sensitive Words: feature extraction is carried out for the preprocessed text data, matching the extracted feature value with the feature value of the sensitive feature database obtained through training; when the matching hits, the system automatically extracts the current sensitive data and its attributes, including the sensitivity level.

In order to improve the recognition rate, Boolean model and probability model are used to represent sensitive data, and CNN multi-layer convolution neural network is used for model training (see the following formula description). The selected sample features include the length of each sensitive data, whether the sensitive data is a data type, the length of the numbers appearing, the length of the letters appearing, and the length of the special symbolsappearing.

*C.* Masking AlgorithmRecommendation

Accordingtothesensitivedataidentifiedbytheabove methods, in different business application scenariosanddatacharacteristics, it is necessary to match themostsuitablemasking algorithm for desensitization. The datato bemasked is used as input, and the best maskingalgorithmisautomatically searched in the selection tree modeljointly.

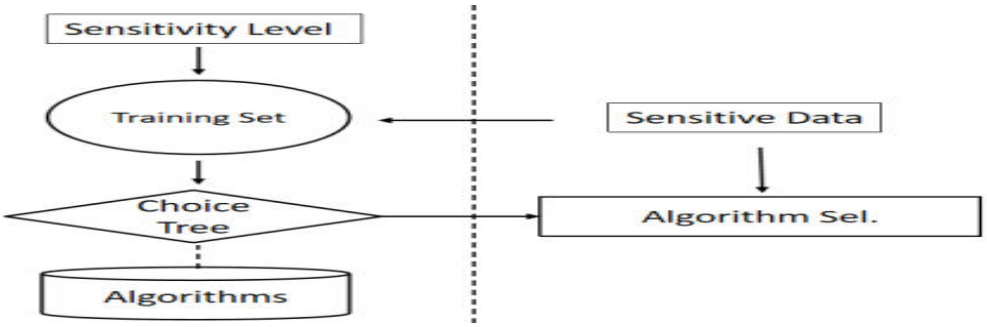


Figure 6. Intelligent Masking Middleware

*D.* Masking AlgorithmExecution

After the masking manager configures the data source, selects the masking scheme and assigns the masking task, the execution of the masking task is started. After the intelligent masking middleware intercepts the data request, it calls a series of data processing procedures, returns the masked data, and records the log of the execution process.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105243 |

## IV. CONCLUSIONS

Aiming at the difficulties of existing government privacy data protection, this paper improves the masking process of the whole life cycle by introducing the artificial intelligence technology such as NLP and machine learning, studies the key problems such as data categorization and classification, intelligent discovery of text sensitive data, automatic recommendation of masking algorithm that need to be solved urgently, and designs and tests the system. The results show that the text sensitive data recognition and algorithmautomatic recommendation significantly improve the masking efficiency, which is a beneficial attempt to promote the systematic, automatic, intelligent and professional government data security protection. The achievements of this work will greatly improve the government's security governance ability in the process of big data sharing and opening up. In the future's work, the further research will focus on intelligent recognition and masking of sensitive data from more unstructured audio, image and video.

## REFERENCES

[1] Chen X Y, Gao Y Z, Tang H L, et al. "Research progress on big data security technology". Sci Sin Inform,2020, 50.
[2] Wang Maolu, et al. "Application of data masking in government data governance and open service", E-GOVERNMENT, May, 2019.
[3] "Governmental Data Guidelines for Categorization and Classification of Data", Sept. 28, 2016.
[4] "Governmental Data Work Instructions for Data Masking", Sept. 28,2016.
[5] Securosis Corporations. "Understanding and Selecting Data Masking Solutions: Creating Secure and Useful Data" [EB/OL]. May 19, 2016.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.379

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details