Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 8841417, 14 pages
https://doi.org/10.1155/2020/8841417

Research Article

Application Research of File Fingerprint Identification Detection Based on a Network Security Protection System

Chunwei Wang,[1,2] **Lina Yu**,[1] **Huixian Chang**,[1] **Sheng Shen**,[1] **Fang Hou**,[1] **and Yingwei Li**[1]

[1]*School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China*
[2]*Beijing Branch, Daqing Oilfield Information Technology Company, Beijing 100043, China*

Correspondence should be addressed to Yingwei Li; lyw@ysu.edu.cn

Received 29 September 2020; Revised 29 October 2020; Accepted 16 November 2020; Published 1 December 2020

Academic Editor: Hongju Cheng

A DLP (data loss prevention) system usually arranges network monitors at the network boundary to perform network traffic capture, file parsing, and strategy matching procedures. Strategy matching is a key process to prevent corporate secret-related documents from leaking. This paper adopts the document fingerprint similarity detection method based on the SimHash principle and customizes the KbS (Keyword-based SimHash) fingerprint, PbS (Paragraph-based SimHash) fingerprint, and SoP (SimHash of Paragraph) fingerprint, three different feature extraction SimHash algorithms for strategy matching to detect. The parsed unstructured data is stored as a file type in.txt format, and then a file fingerprint is generated. Matching the established sensitive document library to calculate the Hamming distance between the fingerprints, the Hamming distance values under different modification degrees are summarized. The experimental results reveal that the hybrid algorithmic strategy matching rules with different levels and accuracy are established. This paper has a reference role for the leakage prevention research of enterprise sensitive data.

1. Introduction

As the enterprise relies increasingly on information systems and information data, information system stability and information data security are directly related to the core competitiveness of enterprises, which puts high demands on enterprise information security and confidentiality [1]. According to the survey, among the ways of leaking sensitive information inside the enterprise, the fact that employees unconsciously outflow sensitive information through the enterprise's home page, mailboxes, instant messaging software, cloud disks, and smartphones connected to a wireless network [2–4] has become an important channel for leaking sensitive information [5, 6]. Therefore, enterprise border networks are particularly important for the timely detection and interception of sensitive data [7, 8].

The previous Information Content Audit platform (1.0) project realized the basic functions of the network monitor and has a good operation effect for the network boundary data capture [9] and document parsing. The identification of sensitive documents is achieved by matching the MD5 fingerprint database with the fingerprint of the document to be detected, but if the document modifies a small amount of content, the MD5 value will become completely different [10, 11]. The leaked data shows that some of the secret-related documents have been modified to a certain extent, where the modification methods include deleting and mixing to form new sensitive documents to avoid the identification of MD5 fingerprints; then, a leak that could not be monitored is formed. The international mainstream research methods for sensitive similar document calculations are based on surface term/word, Vector Space Model (VSM), and hash algorithm [12, 13]. The text similarity method based on the hash algorithm includes the Locality Sensitive Hashing (LSH) algorithm and Locality Preserving Hashing (LPH) algorithm. SimHash is a kind of locality sensitive hashing algorithm, which is ideal for large-scale data processing [14, 15], and it is in line with the research and development mode

of establishing fingerprint database matching sensitive files, which can be directly developed and upgraded on the original platform. At present, there are few research studies on the application of the hash algorithm to the identification of sensitive documents in data loss prevention systems. The use of Hamming distance values to distinguish document similarity lacks effective experimental data [16, 17]. The Information Content Audit platform (2.0) project is an improved project, which includes the improvement of the monitor sensitive matching fingerprint algorithm that will improve the previous single fingerprint algorithm and adopt the hybrid method to achieve similarity calculation to achieve better results [18, 19].

In this paper, the SimHash fingerprint algorithm with three different feature extraction methods is used to study the sensitive document, and the feasibility and effect of multilevel fingerprints formed by mixed multiple hash algorithms in data loss prevention systems are verified, and the monitor function is improved.

The rest of this article is organized as follows. Section 2 introduces the materials and methods of the paper, discussing network monitors and strategy matching, followed by algorithm implementation and experimental scenarios. Section 3 shows the experimental results and discussion. Section 4 describes the fingerprint strategy customization and implementation, and Section 5 summarizes the paper's work and future directions.

2. Materials and Methods

2.1. Network Monitor and Strategy Matching. The main purpose of data leakage prevention products used in enterprises is to monitor network traffic to prevent the transfer of sensitive data to the outside, which usually consists of a central management platform and a monitor, as shown in Figure 1.

It can be seen from Figure 1 that the monitors are arranged in the outlet of the enterprise's external network, which in fact can be used in conjunction with the shunt, and the central management platform is arranged in the enterprise's internal network to manage and configure multiple monitors. The DLP product also provides interfaces to LDAP servers, authentication servers, file servers, and classification and gradation content servers.

DLP products can obtain the specification description of sensitive data from the classification and gradation content servers to establish the determination basis of "sensitivity" that serves as the reference standard for the whole product. The DLP product integrates with the LDAP server to provide information about personnel and departments related to early warning events and provide information for the handling of early warning events [20, 21].

As a core part of the leak prevention, in the enabled state, the monitor continuously captures the traffic on the analysis network [22, 23]. Sensitive data and important traffic elements are monitored through protocols such as SMTP, FTP, and HTTP [24, 25]. This paper is mainly based on the monitor module in the DLP system as the core research content.

Figure 2 shows the functional requirements of content of the monitor, mainly including the packet capture device (fetching the data flow via HTTP, FTP, and SMTP, three agreements), file recognition reader (mainly content parsing for different file formats), and sensitive matcher (sensitive documents matching to identify sensitive documents to block the intercept), three parts.

Among them, the sensitive document matching function that the sensitive matcher needs to complete is an important process of data leakage prevention [26, 27]. How to define enterprise sensitive files, extract the characteristics of sensitive files, and adopt an accurate and efficient algorithm is directly related to the effect of monitor data leakage prevention. According to the characteristics of sensitive documents in this enterprise, this paper uses the SimHash algorithm with three different features to generate corresponding file fingerprints for specified sensitive documents and establish a sensitive document library for storing sensitive files and fingerprint information [28, 29].

From Figure 3, we can see the general flow of the strategy matching fingerprint matching algorithm and the core algorithm used.

The parsed document forms a fixed.txt format. Then, the fingerprint is generated by three algorithms of the fingerprint strategy. Compare fingerprints with fingerprints that exist in the sensitive document libraries; that is, the Hamming distance of the SimHash value is compared to judge whether the strategy hits or not.

2.2. Algorithm Implementation and Experimental Scheme

2.2.1. SimHash and Hamming Distances for Different Feature Extraction. Through the review and research of 1000 enterprise secret-related documents, combined with the specification description of sensitive data collected from the classification and gradation content servers, it is found that most classified documents are enterprise contracts, project-related documents, proxy patents, feasibility study reports, and key technical documents. Moreover, these different versions of documents are widely distributed on the computers of employees involved in document writing, integration, review, modification, and submission, and there are a large number of employees involved, which is prone to leak secrets.

The characteristics of this enterprise on the content level of secret-related documents are as follows. For the article type A, paragraphs are obviously distributed with different length, and long paragraphs are more likely to contain secret-related content. As for the article type B, it is that the distribution of paragraphs is not obvious or the paragraph length is roughly the same and that the nonparagraph forms such as table types also tend to have confidential keywords.

The main idea of the SimHash algorithm is dimension reduction. The high-dimensional feature vector is mapped into an f-bit fingerprint. The Hamming distance of the f-bit fingerprint of the two articles is compared to determine whether the article is repeated or highly approximated. The SimHash algorithm is sophisticated, but it is easy to understand and implement.

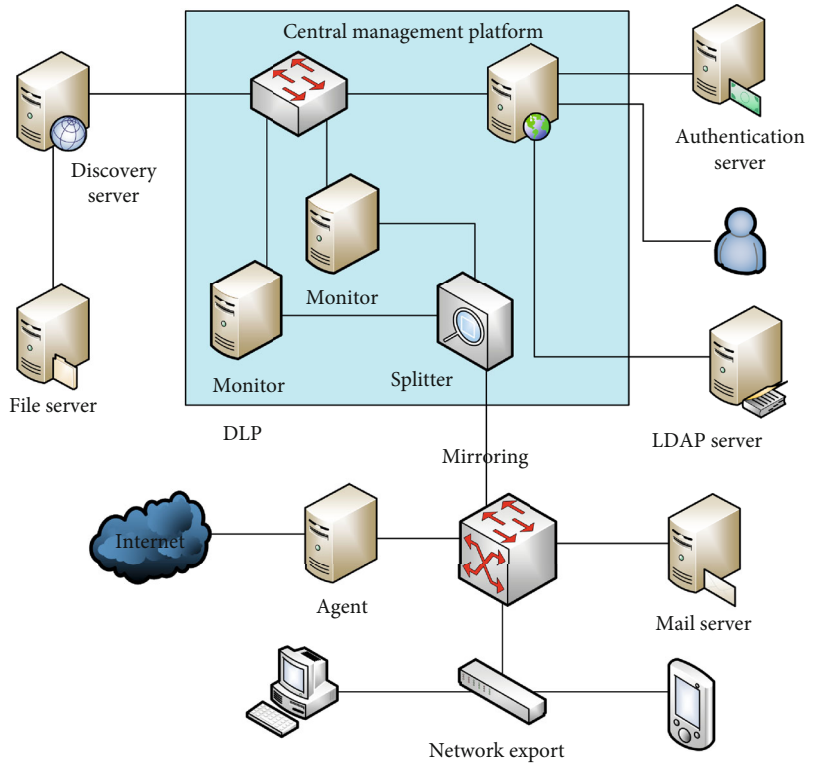

Figure 1: DLP system architecture diagram.

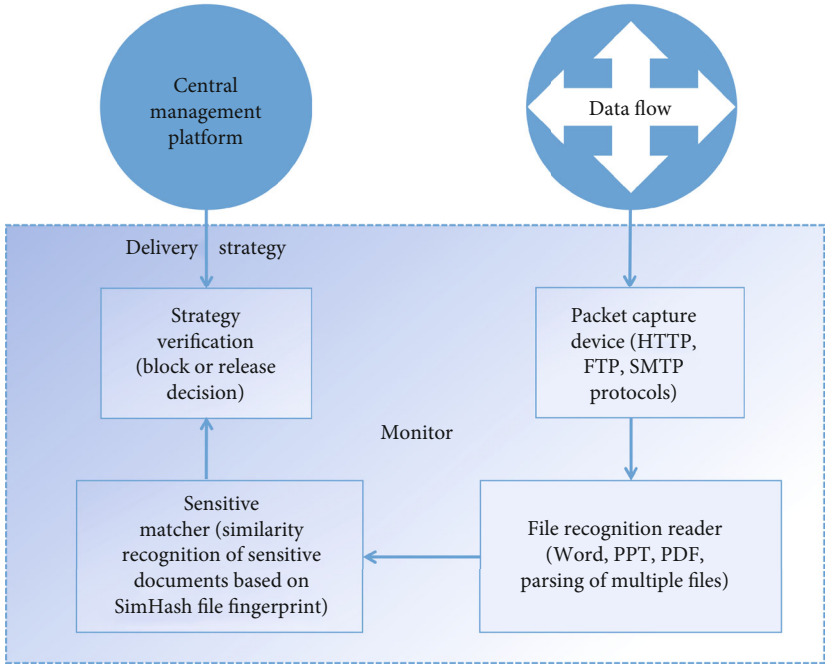

Figure 2: Monitor work flow chart.

In view of the current situation of sensitive document diversity, this paper firstly uses the PbS (Paragraph-based SimHash) fingerprint algorithm and KbS (Keyword-based SimHash) fingerprint algorithm. In order to capture sensitive documents more accurately, the SoP (SimHash of Paragraph) fingerprint algorithm is added [30], jointly implementing the fingerprint strategy. SimHash is used to integrate dynamic and static information to form the features in this paper, and some improvements are made in the feature extraction and weighting process [31]. Figure 4 shows the flow of the three fingerprint algorithms and the detailed process of feature extraction.

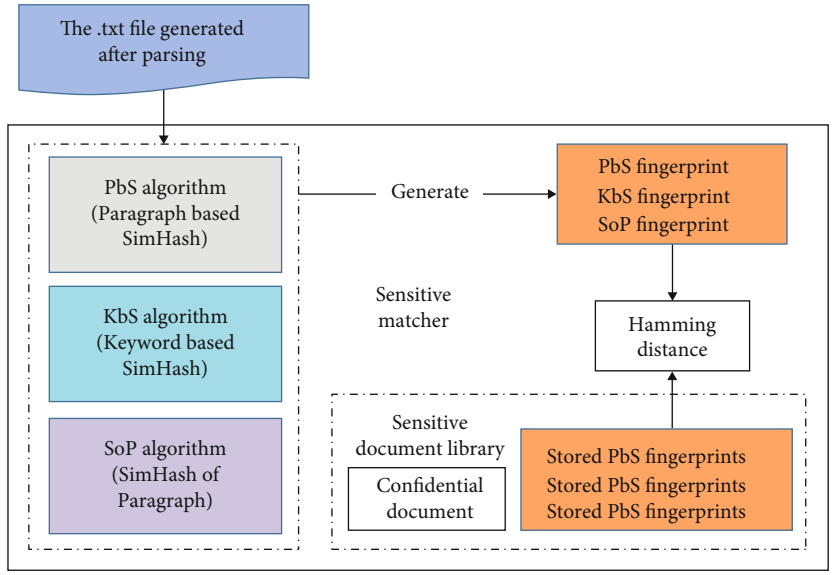

Figure 3: Multiple fingerprint strategy flow chart.

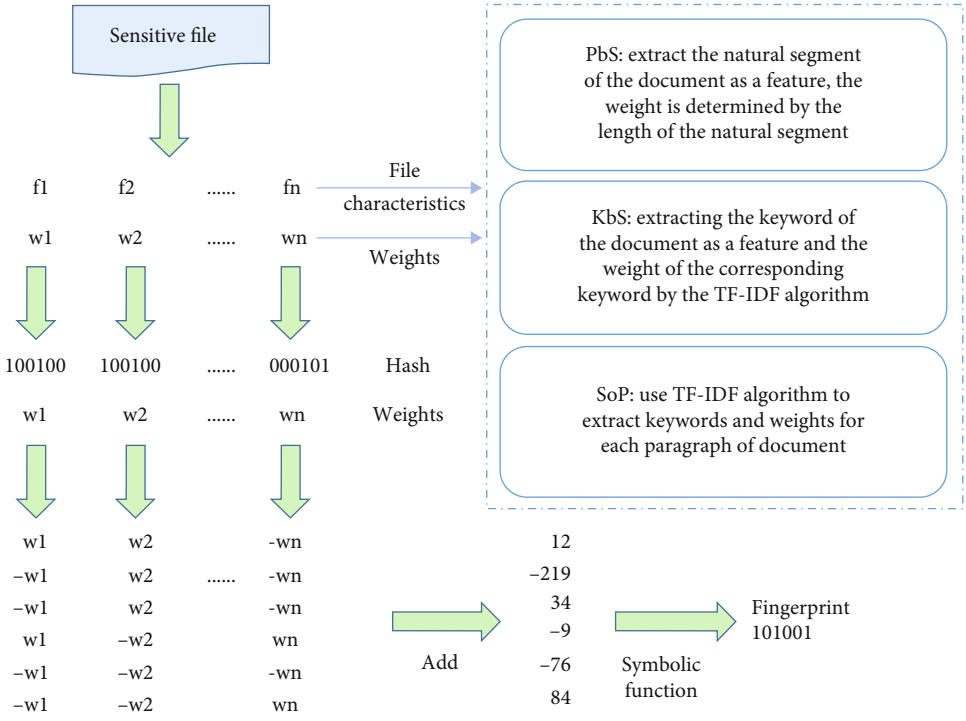

Figure 4: Three SimHash feature extraction methods.

The process of SimHash is roughly divided into the following steps:

(i) *Features*. Sensitive documents processed by the file recognition reader are converted into .txt text, and text features are extracted. PbS uses each paragraph as a characteristic vector; KbS uses the TF-IDF algorithm to extract the first n words of the keyword weight value of the article as features based on the full text; SoP is similar to KbS but extracts n keywords as features on the basis of paragraphs.

(ii) *Weights*. Different document features are extracted, and weight settings are also different. PbS takes the length of characters in each paragraph as the weight value according to the feature that long paragraphs are more likely to appear or represent the main content of the article; KbS uses the TF-IDF algorithm to extract the TF-IDF value of n keywords as the weight value; SoP, similar to KbS, uses the value of TF-IDF calculated by n keywords extracted from each segment as the weight value.

Table 1: Characteristics of three SimHash fingerprint algorithms.

Category	Feature scheme	Weight calculation	Number of fingerprints
PbS	Paragraphs	Natural length	1
KbS	Keywords	TF-IDF value	1
SoP	Keywords	TF-IDF value	N

(iii) *Hash Value*. The hash value of each feature is calculated by the hash function (hash). The hash value is a 64-bit signature composed of the binary numbers 0 and 1.

(iv) *Weighting and Merging*. Multiply the corresponding weight with the hash value, multiply the binary 1 positively and the binary 0 negatively, and get the weighted value of a single feature. Then, the weighted value of each eigenvector is added up to become a sequence string with positive and negative values.

(v) *Dimension Reduction* [32]. For the positive and negative sequence strings, set 1 for each bit if the value is greater than 0, set 0 for each bit if less than or equal to 0, and the resulting 64-bit binary sequence is the final SimHash document fingerprint.

Table 1 shows the characteristics of the three fingerprint algorithms. The keywords of the SoP feature scheme are paragraph keywords; that is, the TF-IDF algorithm is used to extract the keywords of the top n of the current paragraph weight, and the number of fingerprints is the number N of paragraphs of the current document.

For the PbS algorithm, a digital fingerprint is generated for each document, because it only performs hash value processing for each segment, so it is more efficient and saves computing resources. However, if the key paragraph (long paragraph with higher weight) of the sensitive document is changed, it will have a significant impact on the single-paragraph generated hash value and thus affect the fingerprint of the final SimHash. In the case of the KbS fingerprint, the TF-IDF algorithm is integrated to achieve keyword extraction and weight calculation by relatively consuming computing resources in the case of large document length. But for minor changes in sensitive document paragraphs, the keyword extraction and weight of the whole document are almost not affected, and the SimHash fingerprint will not be affected. However, for two documents with similar keywords but different contents, the KbS algorithm will be used in theory to misreport them as sensitive documents, which is less accurate than PbS in this case. Of course, the meaning of data leakage prevention is to prevent leakage of sensitive documents involving confidentiality, and it is better to prevent data leakage with strict levels, so it is safer to integrate multiple different SimHash algorithms.

The SoP fingerprint algorithm is actually a more careful way of getting a better grip on sensitive documents, similar to the KbS principle, except that the objects are changed from full text to paragraphs. Therefore, each paragraph will generate a SimHash digital fingerprint based on TF-IDF to extract keywords and weights. However, the analysis of sensitive documents will be more detailed and accurate. As seen from the number of fingerprints in each document, SimHash fingerprints will be generated for each paragraph of the document, making it easier to see the degree of content changes on the paragraph.

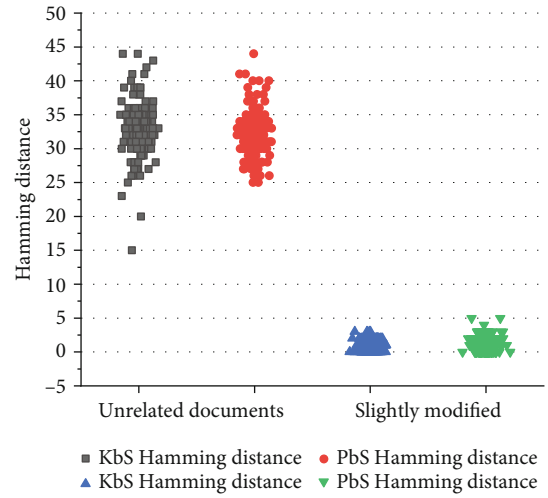

Figure 5: Hamming distance values between unrelated documents and Hamming distance values between different slightly modified documents and original secret-related documents.

With regard to the Hamming distance, the calculation of the Hamming distance is the key to determine whether the document is sensitive or not in this paper. The SimHash fingerprint of the detected document is obtained and compared with the SimHash fingerprint of the secret-related document in the sensitive document library to calculate the Hamming distance. The size of the Hamming distance value determines the similarity between the detected document and the secret-related document. With the increase of the degree of modification, the Hamming distance between the modified document and the original document will also increase. According to the actual situation and experimental test, a reasonable Hamming distance interval is found to determine whether the detection document is sensitive or not.

2.2.2. Fingerprint Strategy Experiment. In order to facilitate the comparison with Hamming distance data between pairs of 20 unrelated documents, this experiment randomly selected 190 secret-related documents in the company. A review of these documents by the Document Security Department found that three types of sensitive documents that are easily leaked are as follows:

(i) Documents formed by minor modification of secret-related documents

(ii) Documents formed by different-scale deletion of secret-related documents

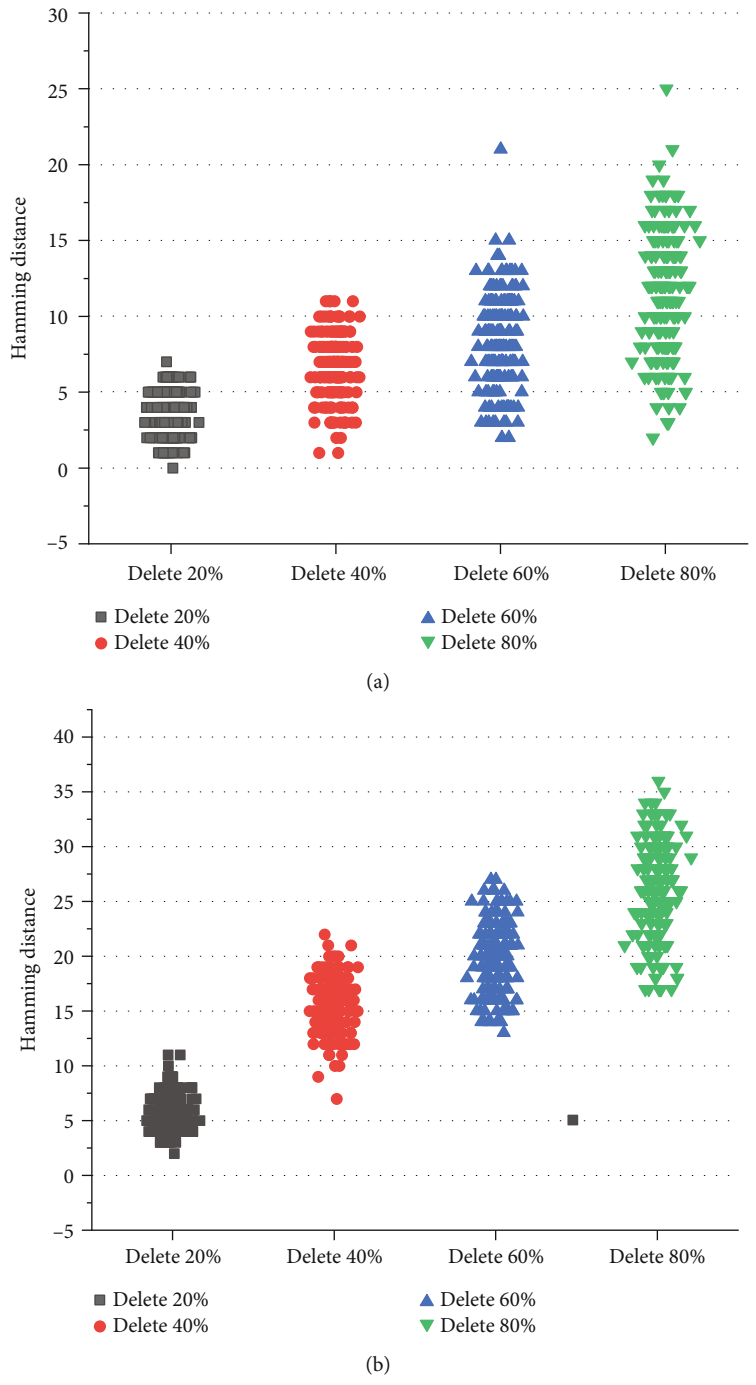

Figure 6: Hamming distance between documents with different degrees of deletion and original secret-related documents. (a) KbS fingerprint Hamming distance. (b) PbS fingerprint Hamming distance.

(iii) Documents formed by different-scale mixing of secret-related documents

For the above three types of sensitive documents, different targeted experiments were conducted: (1) experimental work on the first type of documentation: a small number of artificial modifications will be made to each of the 190 selected articles, such as a small increase in content, a small amount of deletion, and a small number of replacements, and the modifications of the three methods do not exceed 10% to simulate the formation of a revised version of a sensitive document; (2) experimental work on the second type of documentation: delete 20%, 40%, 60%, and 80% of the original secret-related documents to simulate that the document to be detected has only a small portion of the secret-related document; and (3) experimental work on the third type of documentation: delete 20%, 40%, 60%, and 80% of the total content of the original document, and then add 20%, 40%,

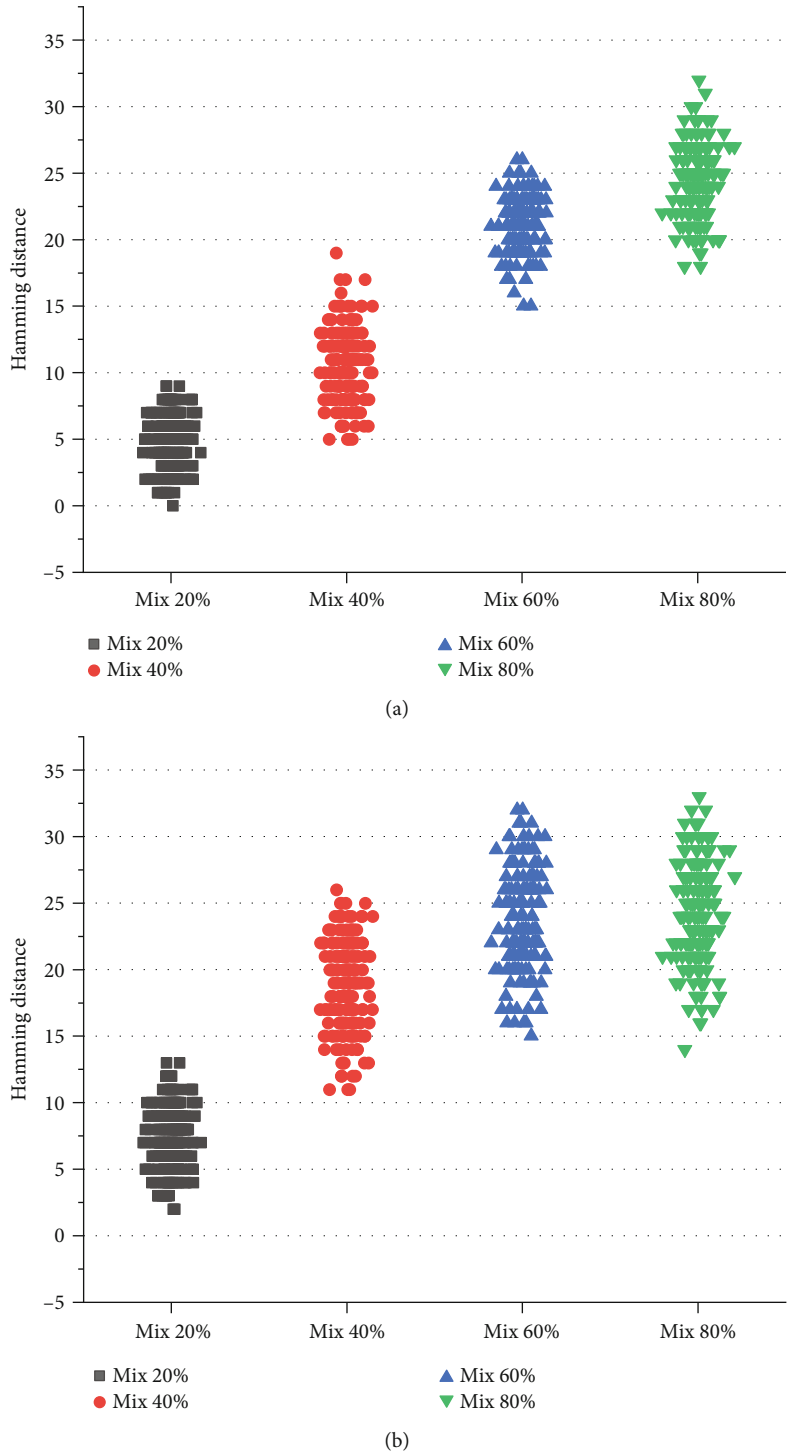

Figure 7: Hamming distance between different mixed degree documents and original secret-related documents. (a) KbS fingerprint Hamming distance. (b) PbS fingerprint Hamming distance.

60%, and 80% irrelevant content to simulate different levels of mixed documents. At the same time, 20 unrelated documents are prepared, and the Hamming distance between them is calculated. 190 Hamming distance values are obtained to explore the Hamming distance range between unrelated documents, so as to distinguish the Hamming distance between secret-related documents and sensitive documents.

3. Results and Discussion

3.1. KbS and PbS Fingerprint Experiment Results. As shown in Figure 5, the Hamming distance value between 20 unrelated documents, regardless of the PbS fingerprint or KbS fingerprint, is concentrated above 26, and only individual values are lower than 26. The Hamming distance values between the

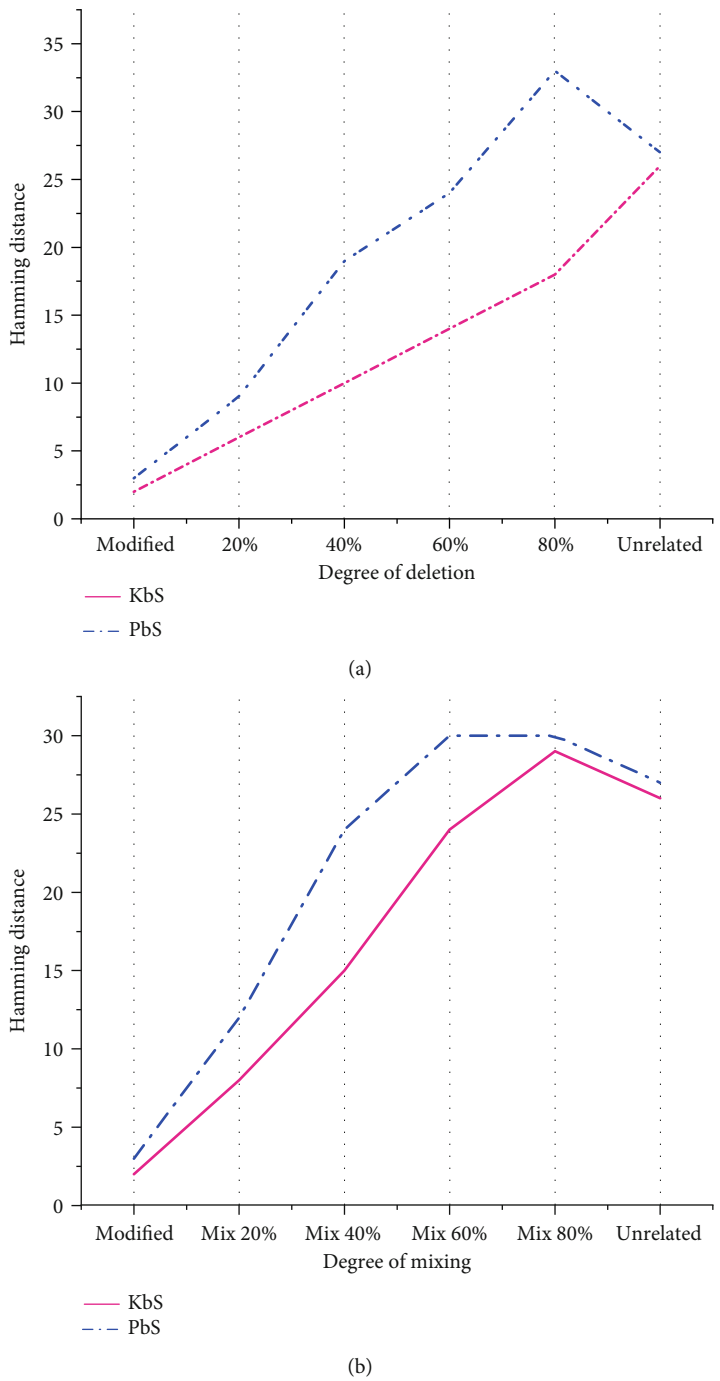

Figure 8: Hamming distance value between the documents under different modified situations and the original confidential documents. (a) The maximum Hamming distance between KbS and PbS under different deletion degrees. (b) The maximum Hamming distance between KbS and PbS under different mixing degrees.

sensitive documents and the confidential documents that are slightly modified to form different versions are concentrated between 0 and 3.

For the experimental results of different degrees of deletion shown in Figure 6, the Hamming distance between the PbS fingerprint and the KbS fingerprint and the original secret-related document will increase with the increase of the degree of deletion. In terms of the degree of discrimination of different degrees of deletion, the degree of Hamming distance distinction between KbS fingerprints is not obvious, and the levels are mixed, while the PbS fingerprints are relatively obvious and hierarchical. Considering the overall amplitude, the KbS fingerprint Hamming distance value is relatively small and concentrated, but the PbS fingerprint Hamming distance is relatively large and scattered. In view of the amplitude changes caused by different degrees of deletion, the variation of the KbS fingerprint Hamming distance is not large, while the PbS fingerprint changes are relatively

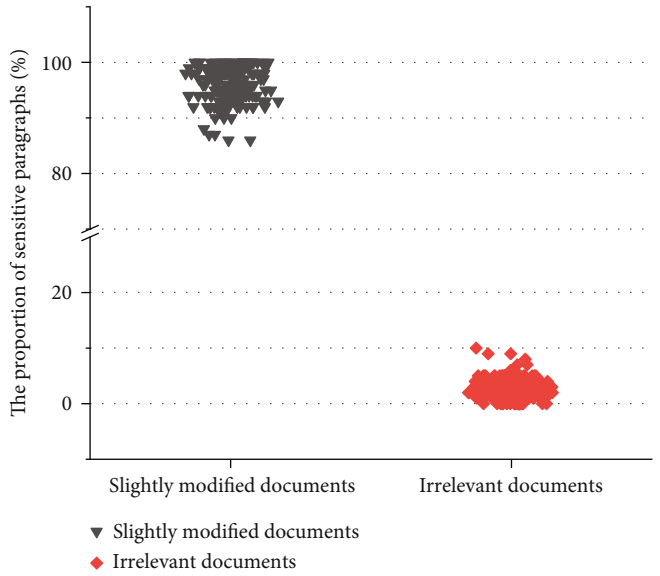

Figure 9: Proportion of sensitive natural segments in the documents.

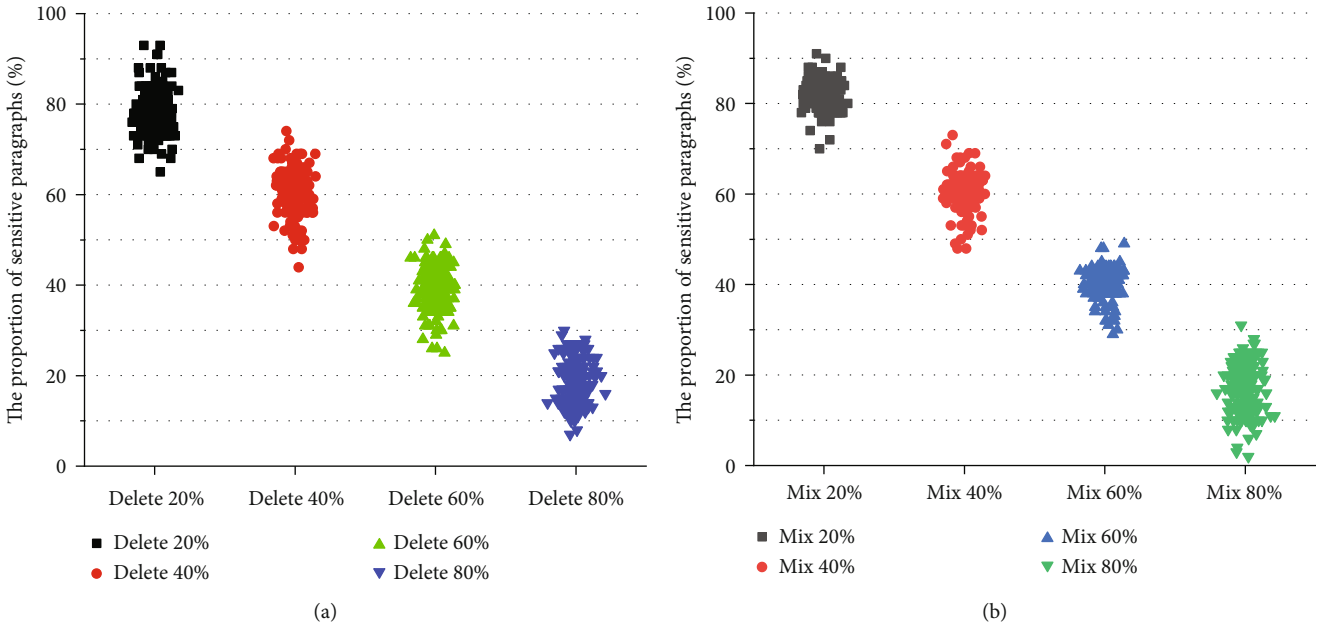

Figure 10: Proportion of sensitive natural segments in the documents under different modification situations. (a) Proportion of sensitive natural segments in documents at different levels of deletion. (b) Proportion of sensitive natural segments in the documents at different levels of mixing.

large. The experimental results under different mixing conditions are shown in Figure 7. KbS fingerprints increase with the increase of the degree of mixing, and Hamming distance increases at the same time. When the degree of mixing exceeds 60%, the variation of the Hamming distance value decreases. Most of them are concentrated in the interval between 18 and 28. When the degree of mixing of PbS fingerprints is greater than 20%, the variation of Hamming distance values decreases which are mostly concentrated between 15 and 30 and are more scattered.

The changes of the Hamming distance are clearly shown in Figure 8, in which the Hamming distance value of the documents under different degrees of modification is the maximum Hamming distance value after removing a few top values and bottom values (2%), while the Hamming distance value of irrelevant documents is the minimum Hamming distance value after removing a few bottom values (2%). If the maximum Hamming distance of the changed documents does not exceed the minimum value of irrelevant documents, it means that a better policy matching effect is generated and sensitive documents are captured accurately.

3.2. SoP Fingerprint Experiment Results. For the experimental implementation of the SoP fingerprint algorithm, the

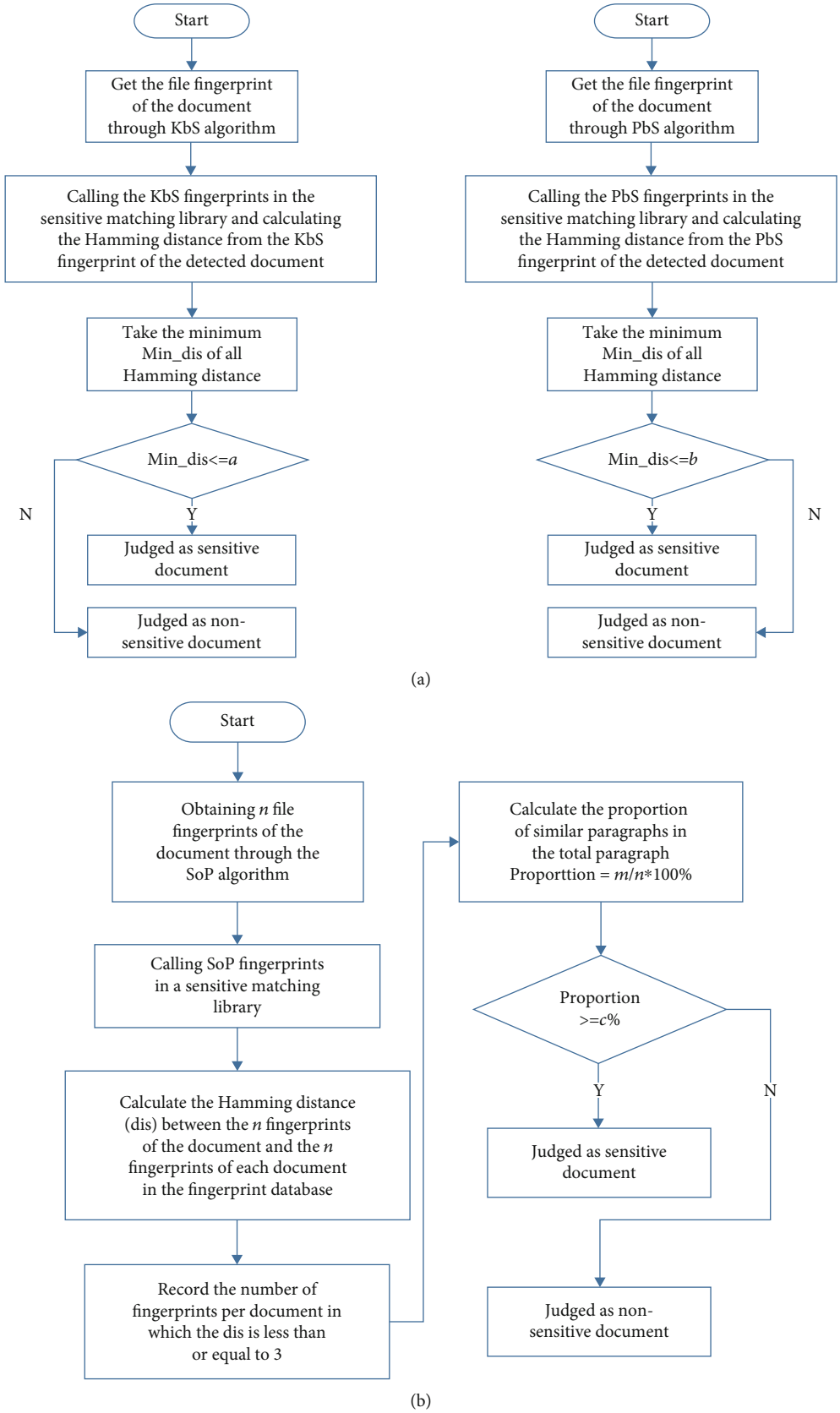

Figure 11: Process flow corresponding to different fingerprint algorithms. (a) KbS fingerprint and PbS fingerprint processing flow chart. (b) SoP fingerprint processing flow chart.

Hamming distance test between three common sensitive documents and original documents was conducted as same as the other two algorithms, which includes the sensitive documents under different modifications, different degrees of content deletion, and different degrees of content mixing. Due to the nature of SoP fingerprints that there is an unfixed number of fingerprints per document, SoP fingerprints no longer use the Hamming distance between documents as the basis for judging whether they are sensitive documents. When the Hamming distance between two documents is less than or equal to 3, the content between the documents is extremely similar. Therefore, in this experiment, the proportion of the number of SoP fingerprints in the document with a Hamming distance less than or equal to 3 to the total number of SoP fingerprints in the entire document is used as the criterion to determine whether the document is a sensitive document.

As shown in Figure 9, a number of SoP fingerprints in each of the 20 irrelevant documents are compared in pairs. The number of Hamming distances less than or equal to 3 accounts for the proportion of the total number of SoP fingerprints in the documents between 0% and 5%. Individual values are higher than 5%. The proportion of sensitive natural segments of the slightly modified document is concentrated above 90%, and only individual values are lower than 90%.

For algorithm fingerprints of SoP, test experiments with different levels of deletion and different levels of mixing were also carried out. As shown in Figure 10, for two different modification situations, the proportion of sensitive natural segments in the document decreases as the degree of modification increases. Since the extraction of paragraph keywords is affected by the amount of paragraph content, individual values will appear relatively discrete. When the degree of modification does not exceed 40%, the proportion of sensitive natural segments in the document is basically higher than 50%. When the degree of modification exceeds 60%, the proportion of sensitive natural segments in the document is basically lower than 50%. When the degree of modification exceeds 80%, the proportion of sensitive natural segments in the document is mostly higher than 10%, and only the individual values are lower.

4. Fingerprint Strategy Customization and Implementation

According to the different needs of different periods in the enterprise, the process of strategy customization will be different. It is roughly divided into the following three typical demand periods. Period 1 is the period of the daily operation of the enterprise, and the level of data loss prevention is low during this period, so the detected document is judged to be a sensitive document when it is very similar to the secret-related document and contains most of the content. In period 2, it is about the company issues patents, software copyrights, etc. The requirement of data loss prevention for this period is medium, so the detected documents are judged to be sensitive documents when they are similar to the secret-related documents and contain some content. In period 3, when the enterprise has a large-scale activity of the "protective network action" or the introduction of important project documents, this period requires a higher level of data loss prevention, so for the detected document, as long as it contains any paragraph content of the secret-related document, it will be recognized as a sensitive document.

Table 2: Parameter values in different periods.

Different periods	Parameter a	Parameter b	Parameter c
1	3	3	90%
2	14	24	50%
3	26	27	10%

According to the three fingerprint algorithms, three different design flows are adopted, as shown in Figure 11. Parameters are the key to judge whether documents in different periods are involved in confidentiality, so they need to be set according to the above experimental Hamming distance distribution. The values of the parameters a, b, and c determine the strictness of the strategy customization, in other words, the level of the data loss prevention.

From the previous chapter, according to the analysis of the experimental data of the Hamming distance, Figure 8 clearly shows that when the Hamming distance between documents is less than or equal to 3, the detected document is a modified version of the secret-related document. The ratio of sensitive segments in slightly modified documents is concentrated at more than 90%, and the similarity is extremely high, which meets the needs of the first period. Therefore, the parameters a and b in Figure 11 can be set to 3, and the parameter c is set to 90%; that is, similar paragraphs account for 90% of the total paragraphs and the documents are considered to be sensitive.

It can be seen from Figure 8 that when the Hamming distance between documents is between 3 and 24, the detected documents are mixed with different levels of secret-related content, meeting the needs of the second period, so a in Figure 11 is set to 15, b is set to 24, and the parameter c is set to 50%. That is, similar paragraphs accounting for more than 50% of the total paragraphs are considered sensitive documents. For the requirements of period 3, any paragraph content containing the secret-related document in the detected document will be judged as a sensitive document. At this time, the accuracy of the KbS fingerprint and PbS fingerprint is insufficient. However, to facilitate the comparison of experimental data, the experimental parameters of KbS and PbS are set as the minimum value of the Hamming distance between unrelated documents; that is, a and b were set as 26 and 27, respectively. Figure 10 shows that when the degree of modification exceeds 80%, the sensitive segment ratio is mostly higher than 10%; thus, the value of c is set to 10%. That is, similar paragraphs accounting for more than ten percent of the total paragraphs are identified as sensitive documents. Table 2 shows the parameter values set in different periods.

In the strategy customization process, it is necessary to consider the accuracy and efficiency of the algorithm selected in different demand periods. Figure 12 shows the time spent

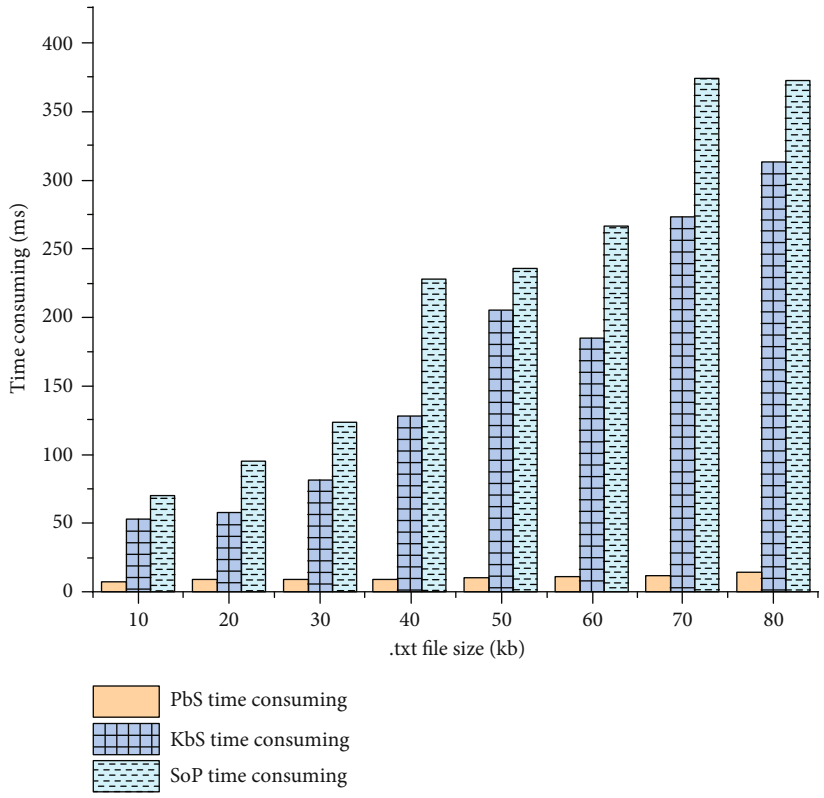

Figure 12: Fingerprint time-consuming graph of three algorithms.

on fingerprint calculation under different sizes of the parsed .txt file. After strategy matching, the files that have been judged as sensitive documents need to be reviewed by the Safety Manager again for human judgment, and the purpose of strategy optimization is to formulate a more reasonable strategy to reduce labor. Therefore, the text of this test is reviewed by the Safety Manager. The Safety Manager judged the confidentiality of the text under different levels of demand according to the needs of different periods. In the enterprise intranet, 1000 documents are randomly selected as test texts, which contain sensitive documents involving different degrees and irrelevant documents. The results of the Safety Manager review of the documents at different levels are as follows: at the low level, 87 documents were identified as sensitive documents, with 238 in the middle level and 385 in the high level.

In order to verify the accuracy of identifying sensitive documents between different algorithms, three indicators of Equations (1), (2), and (3) are used:

$$\text{Precision} = \frac{\text{Num}_{\text{correct}}}{\text{Num}_{\text{total}}}, \tag{1}$$

$$\text{Recall} = \frac{\text{Num}_{\text{correct}}}{\text{Num}_{\text{actual}}}, \tag{2}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{3}$$

Table 3: Accuracy indicators of three fingerprints under different levels.

Levels	Fingerprint process	Precision (%)	Recall (%)	F1 (%)
Low	KbS	80.95	97.7	88.53
	PbS	100	95.4	97.64
	SoP	97.7	97.7	97.7
Medium	KbS	87.73	96.21	91.78
	PbS	74.55	88.65	80.99
	SoP	85.33	95.37	90.07
High	KbS	58.91	81.56	68.41
	PbS	47.52	91.95	62.69
	SoP	91.37	96.36	93.8

Among them, $\text{Num}_{\text{correct}}$ is the number of sensitive documents identified correctly, $\text{Num}_{\text{total}}$ is the number of sensitive documents identified, and $\text{Num}_{\text{actual}}$ is the total number of sensitive documents. This method is used to test the above 1000 documents. Table 3 shows the accuracy indicators of each fingerprint process at different levels. From Table 3, it can be seen that in the period when the data loss prevention level is low, the PbS fingerprint processing flow and the SoP fingerprint processing flow have higher F1 values, but the KbS fingerprint processing flow F1 value is relatively low. Combined with the time consumed of the algorithm of Figure 12 to comprehensive consideration, the efficiency of PbS has an absolute advantage, so it is most appropriate to adopt the PbS fingerprint processing process in this case. In

the period when the data loss prevention level is medium, the F1 values of the KbS fingerprint processing flow and the SoP fingerprint processing flow are above 90%, and there is a high recall rate. Combined with the processing efficiency problem, the KbS fingerprint processing flow has advantages, so in this case, the KbS fingerprint processing process is the most suitable.

However, in period 3 with high requirements for data leakage prevention, the F1 value of the KbS fingerprint and PbS fingerprint processing process is relatively low, while the F1 value of the SoP fingerprint processing process is above 90%, and SoP has high precision and recall rates. Therefore, SoP fingerprint processing can only be adopted in this situation, according to the data in Table 3.

5. Conclusion

In this paper, three kinds of fingerprint algorithms are combined with the targeted secret-related documents in the enterprise for experimental analysis. It provides experimental reference and data reference for the relationship between sensitive documents and Hamming distances between secret-related documents in different situations. It also provides the basis for selecting the parameters of the fingerprint strategy in this paper and calculates and tests the efficiency of three different fingerprints. The fingerprint strategy of this system is formulated for three special periods in the enterprise. Using different fingerprint strategies at different periods can more accurately detect sensitive documents, but occasionally there are cases of false positives. In future work, the system can be optimized and more detailed exploration can be carried out on the basis of the fingerprint strategy experiment, such as the influence of more detailed document changes on the Hamming distance, so as to meet the requirements of a more strict or detailed leak prevention strategy, making the leak prevention system put into production more efficiently and accurately.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation (61827811), National Defense Basic Research Program (JCKY2019407C002), Hebei Provincial Education Departments Support Plan (SLRC2019042), Hebei Province Funding Project for the Introduction of Overseas Students (C20200364), and China National Petroleum Corporation Information Technology Construction Project (CNPC-IT-2018-N001).

References

[1] H. Liang, A. Xian, M. Mao, P. Ni, and H. Wu, "A research on remote fracturing monitoring and decision-making method supporting smart city," *Sustainable Cities and Society*, vol. 62, 2020.

[2] H. Cheng, N. Xiong, A. V. Vasilakos, L. Tianruo Yang, G. Chen, and X. Zhuang, "Nodes organization for channel assignment with topology preservation in multi-radio wireless mesh networks," *Ad Hoc Networks*, vol. 10, pp. 760–773, 2012.

[3] W. Guo, N. Xiong, H. Chao, S. Hussain, and G. Chen, "Design and analysis of self-adapted task scheduling strategies in wireless sensor networks," *Sensors*, vol. 11, pp. 6533–6554, 2011.

[4] N. Xiong, A. Vandenberg, and W. Han, "Green cloud computing schemes based on networks: a survey," *IET Communications*, vol. 6, pp. 3294–3300, 2012.

[5] I. You, M. R. Ogiela, I. Woungang, and K. Yim, "Innovative security technologies against insider threats and data leakage," *International Journal of Computer Mathematics*, vol. 93, pp. 236–238, 2014.

[6] X. Wang, J. Shi, and L. Guo, "Towards analyzing traceability of data leakage by malicious insiders," *Trustworthy Computing and Services*, 2012.

[7] S. Alneyadi, E. Sithirasenan, and V. Muthukkumarasamy, "A survey on data leakage prevention systems," *Journal of Network and Computer Applications*, vol. 62, pp. 137–152, 2016.

[8] A. Shrivastava and P. Li, "In defense of MinHash over Simhash," *Eprint Arxiv, arXiv*, vol. 1407, 2014https://arxiv.org/abs/1407.4416.

[9] H. Zheng, W. Guo, and N. Xiong, "A kernel-based compressive sensing approach for mobile data gathering in wireless sensor network systems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 48, pp. 2315–2327, 2017.

[10] J. Lin, Q. Huang, and J. Zhang, "Method of data tamper detection by using improved MD5 algorithm," *Computer Engineering & Applications*, vol. 44, pp. 148–150, 2008.

[11] Q. Lv, F. Duan, Y. Wu, and J. He, "Similarity retrieval algorithm based on multilevel fingerprint comparison matrix," *International Symposium on Communication Engineering & Computer Science (CECS 2018)*, vol. 2018, 2018.

[12] L. Xu, S. Sun, and Q. Wang, "Text similarity algorithm based on semantic vector space model," , 2016*IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, vol. 2016, 2016.

[13] R. Gula and M. Ranum, "System and method for using file hashes to track data leakage and document propagation in a network," 2012, Patent ID US 13/403,108,.

[14] Y. Li, F. Liu, Z. Du, and D. Zhang, "A Simhash-based integrative features extraction algorithm for malware detection," *Algorithms*, vol. 11, p. 124, 2018.

[15] Y. Zheng, Z. Xie, Y. Li, G. Shen, and H. Liu, "Spatial vibration of rolling mills," *Journal of Materials Processing Technology*, vol. 213, pp. 581–588, 2013.

[16] Z. Huang, J. Tang, G. Shan, J. Ni, Y. Chen, and C. Wang, "An efficient passenger-hunting recommendation framework with multitask deep learning," *IEEE Internet of Things Journal*, vol. 6, pp. 7713–7721, 2019.

[17] R. A. F. Alvarenga, J. Dewulf, H. Van Langenhove, and M. A. J. Huijbregts, "Exergy-based accounting for land as a natural resource in life cycle assessment," *The International Journal of Life Cycle Assessment*, vol. 18, pp. 939–947, 2013.

[18] H. Wang, S. Liu, and Z. Jia, "A fingerprint of paragraph generation approach for detecting similar document," *Information Technology Journal*, vol. 13, pp. 2309–2317, 2014.

[19] M. Mekonnen, T. Sewunet, M. Gebeyehu, B. Azene, and A. M. Melesse, "GIS and remote sensing-based forest resource assessment, quantification, and mapping in Amhara region, Ethiopia," *Landscape Dynamics, Soils and Hydrological Processes in Varied Climates*, pp. 9–29, 2016.

[20] S. Mellino and S. Ulgiati, "Monitoring regional land use and land cover changes in support of an environmentally sound resource management," *Sustainable Development, Knowledge Society and Smart Future Manufacturing Technologies*, pp. 309–321, 2015.

[21] S. Shimada, T. Ohsawa, T. Kogaki, G. Steinfeld, and D. Heinemann, "Effects of sea surface temperature accuracy on offshore wind resource assessment using a mesoscale model," *Wind Energy*, vol. 18, pp. 1839–1854, 2015.

[22] H. Cheng, Z. Su, N. Xiong, and Y. Xiao, "Energy-efficient node scheduling algorithms for wireless sensor networks using Markov random field model," *Information Sciences*, vol. 329, article S0020025515006945, pp. 461–477, 2016.

[23] H. Cheng, Z. Xie, L. Wu, Z. Yu, and R. Li, "Data prediction model in wireless sensor networks based on bidirectional LSTM," *EURASIP Journal on Wireless Communications and Networking*, vol. 203, 12 pages, 2019.

[24] J. Lyons, A. Dehzangi, R. Heffernan et al., "Predicting backbone Cα angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network," *Journal of Computational Chemistry*, vol. 35, pp. 2040–2046, 2014.

[25] L. Ferrer, Y. Lei, M. Mclaren, and N. Scheffer, "Study of senone-based deep neural network approaches for spoken language recognition," *IEEE/ACM Transactions on Audio Speech & Language Processing*, vol. 24, pp. 105–116, 2016.

[26] H. Liang, J. Zou, Z. Li, M. J. Khan, and Y. Lu, "Dynamic evaluation of drilling leakage risk based on fuzzy theory and PSO-SVR algorithm," *Future Generation Computer Systems*, vol. 95, pp. 454–466, 2019.

[27] H. Liang, J. Zou, K. Zuo, and M. J. Khan, "An improved genetic algorithm optimization fuzzy controller applied to the wellhead back pressure control system," *Mechanical Systems and Signal Processing*, vol. 142, Article ID 106708, 2020.

[28] A. Ardakani, F. Leduc-Primeau, N. Onizawa, T. Hanyu, and W. J. Gross, "VLSI implementation of deep neural network using integral stochastic computing," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 25, pp. 2688–2699, 2017.

[29] M. J. Kang and J. W. Kang, "Intrusion detection system using deep neural network for in-vehicle network security," *PLOS ONE*, vol. 11, no. 6, article e0155781, 2016.

[30] Y. Miao, H. Zhang, and F. Metze, "Speaker adaptive training of deep neural network acoustic models using I-vectors," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 23, pp. 1938–1949, 2015.

[31] M. Kolbk, Z. H. Tan, J. Jensen et al., "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE/ACM Trans Audio Speech Lang Process*, vol. 25, no. 1, pp. 153–167, 2017.

[32] Y. Zhang, R. Zhu, Z. Chen, J. Gao, and D. Xia, "Evaluating and selecting features via information theoretic lower bounds of feature inner correlations for high-dimensional data," *European Journal of Operational Research*, 2020.