A Survey on Privacy-Preserving Data Mining Methods

Yunlu Yang*, Yajian Zhou*
School of Cyberspace Security. Beijing University of Posts and Telecommunications. Beijing, China

*Corresponding author e-mail: yajian@bupt.edu.cn, *yangyunlu@bupt.edu.cn

Abstract. In recent years, the increase in massive data in various fields has promoted the development of data mining. Still, the storage and mining of user data bring about the threat of privacy leakage. Researches on privacy-preserving data mining have become an increasingly significant research area. Focusing on the hidden dangers of privacy leakage in data mining methods, three leading privacy-preserving technologies came into being, which are data distortion technology, data encryption technology, and restricted publication technology. In this paper, we introduce and summarize the latest researches based on these privacy-preserving technologies. Besides, we describe the state-of-art research trends on privacy-preserving data mining methods in image processing and natural language processing.

1. Introduction
The rapid evolution of big data technologies and artificial intelligence has a double-edged effect. On the one hand, it becomes more accessible and more convenient to collect, store, and publish massive amounts of data (e.g., image, text, audio), which makes incredible achievements in all walks of life. In the foreseeable future, big data analytics will continue to play a more and more significant role in promoting the development of the economy and society [1].

On the other hand, rich semantics contained within the exchanged or released data make it easy for attackers to extract and synthesize the privacy information of individuals [2]. As a result, data protection, especially privacy protection, becomes daunting. It suggests that privacy preservation becomes an arduous job and attracts more and more attention from both researchers, enterprises, and decision-makers of governments during the past decade. Therefore, designing a reliable privacy-preserving mechanism, without leading any unaccepted compromise to the performance of big data analytics, remains a promising topic during recent years [3].

To tackle a series of security threats and challenges in big data and artificial intelligence, researchers have proposed many countermeasures and solutions. According to the data life cycle, the big data privacy-preserving technologies can be divided into four categories: privacy-preserving data publication, privacy-preserving in data storage, privacy-preserving data mining, and privacy-preserving in data usage [4]. In 2000, the concept of privacy-preserving data mining (PPDM) was firstly proposed by Agrawal [5]. PPDM not only provides an order of magnitude enhancement in mining efficiency but also ensures the security of sensitive data or delicate patterns. On this basis, scholars have conducted various researches, combining privacy-preserving technologies with existing data mining methods, such as image processing, natural language processing.
In this paper, we initially introduce and summarize the latest privacy-preserving technologies, including data distortion technology, data encryption technology, and restricted publication technology. Subsequently, we conclude the research trends for data mining methods based on privacy-preserving in image processing and natural language processing.

2. Research Progress of Privacy-Preserving Technology

2.1. Data Distortion Technology
Data distortion technology performs privacy-preserving by disturbing the original data [6]. The mainstream algorithms of data distortion technology include randomization, blocking and cohesion, and differential privacy [7]. The perturbed data has two main characteristics:

(1) The attacker cannot discover or infer the real original data. In other words, the attacker cannot reconstruct the original data from the released data.

(2) The statistical significance or primary correlation mode of data is preserved, and the analyzed results on the distorted data are close to the results obtained from the original data analysis.

Comparing with other methods based on data distortion principle, differential privacy throws light on the solution to privacy preservation. The outstanding feature of differential privacy technology makes it high universal and interpretable [8]. Differential privacy optimizes the effect of privacy protection by adding less noise, but without affecting the analysis results of the overall data [9]. The privacy budget is a crucial parameter of differential privacy, which controls the amount of added noise and provides a measure of privacy protection performance. As the added noise increases, the distortion of the data becomes more serious, and the privacy protection effect will be better. In the process of data analysis, the balance between the availability of data analysis results and the degree of privacy protection can be achieved through the adjustment of parameters [10].

Data mining methods based on differential privacy can be divided into two categories according to different interaction methods [11]. The first category is an interactive framework. In this framework, the data miner submits query statements to the data owner, and the data owner returns the query results to the data miner after adding noise. The second category is a non-interactive framework. The two implementations of the non-interactive framework are as follows:

(1) The data owner preprocesses the data based on the differential privacy mechanism before the data is released and then sends the processed data to the data miner.

(2) The data miner submits the data mining algorithm to the data owner. The data owner calculates the appropriate privacy budget, and then brings the privacy budget into the calculation process, finally returns the analysis result to the data miner.

Differential privacy technology has many applications in the field of data mining. Table 1 introduces some researches on differential privacy protection technology. Applications of differential privacy in data mining
| Method Name          | Solution description                                                                 | Advantage                                                                 | Disadvantage                                                  |
|---------------------|---------------------------------------------------------------------------------------|---------------------------------------------------------------------------|---------------------------------------------------------------|
| DiffID3             | Combining the index mechanism to generate an ID3 decision tree for data analysis      | Simple implementation and high classification accuracy                     | Controlling privacy budget parameter assignments is difficult |
| DP-Bayes            | Adding Noise to Parameters in Naive Bayes Classifier for Differential Privacy          | The simple structure of the model                                          | Relying on assumptions about data distribution, and the smaller the data set, the higher the noise |
| Output perturbation mechanism | After training the model, adding noise to the model for differential privacy    | High applicability and classification accuracy in smaller data sets         | SVM coefficients can cause significant errors, making it challenging to process more massive data sets |
| SuLQ-K-Means        | Posting estimated values for cluster centroids and the number of records              | Ability to protect cluster centroid position                                | The amount of noise calculated by sensitivity is large, which reduces the availability of clustering results |
| Distributed training method | Injecting noise into "gradients" of parameters to protect privacy in neural networks | Ability to withstand poison attacks                                         | Consuming an unnecessary privacy budget                       |
| Moments Accountant  | Tracking privacy spending and enforcing appropriate privacy protections                | Ability to adequately protect sensitive information in the training set     | The privacy budget is accumulated in proportion to the number of training periods and the number of shared parameters |

2.2. Data Encryption Technology

Data encryption technology uses encryption algorithms and encryption keys to convert plain text into ciphertext, and introduces the data encryption mechanism into interactive computing protocols to achieve secure calculation of data confidentiality. The most representative technology is homomorphic encryption. In 1978, homomorphic encryption was first proposed by Rivest and applied in the bank [12]. The result of an algebraic operation on the ciphertext through homomorphic encryption technology is still encrypted. Besides, the same results as those obtained by performing the same algebra operation on the plaintext can be obtained after decryption.

Before submitting data to the data processing center, the bank first encrypts the data with some homomorphic encryption algorithms. Then, the data processing center uses the data mining model to analyze and process the encrypted data. Finally, the data processing center returns the encrypted data to the bank, and the bank uses its private key to decrypt it to get the corresponding result. In this process, the data processing center only obtains the encrypted data, and also performs data mining based on the encrypted data, so the bank's data must be secure. However, data encryption and decryption will consume large computing resources, resulting in homomorphic encryption technology being more applicable in distributed systems.
The research status of data mining methods based on homomorphic encryption technology is as follows.

2.2.1. "Noise" Optimization of Homomorphic Encryption Technology. Focusing on excessive noise introduced by homomorphic encryption, Graepel proposed a scheme that mathematically expressed the prediction function of the model as a low-order polynomial [13]. When training the data mining model, the scheme can not only effectively reduce the number of homomorphic operations on encrypted data, but also limit the operation mode to addition homomorphism and multiplication homomorphism. By applying to Fisher linear discriminator, the scheme has achieved practical results.

2.2.2. Homomorphic encryption for large-scale data sets. The explosive increase of data on the Internet has caused an enormous impact on traditional personal privacy protection issues. In order to deal with the privacy protection of large-scale data sets and high-dimension data, Wu designed a scheme of homomorphic encryption for large-scale data sets [14]. Previously, the homomorphic operation was implemented using batch calculation and CRT-based message encoding technology. Subsequently, a statistical analysis, such as linear regression, was performed on the encrypted data. This scheme is suitable for scenarios with multiple sources of data.

2.2.3. The combination of homomorphic encryption and deep neural networks. Based on the Chebyshev approximation theory, Dowlin replaced the non-linear activation function of the neural network model with a low-order polynomial function and designed a CryptoNets which can process ciphertext [15]. Experiments on the MNIST data set illustrate the rationality of the model. Chabanne found that the performance of CryptoNets on deep neural networks was unsatisfactory sometimes, and proposed a convolutional neural network model with higher accuracy than CryptoNets [16]. The proposed model firstly successfully combined homomorphic encryption with deep neural networks. Besides, Aono proposed a logistic regression model that can calculate ciphertext using additively homomorphic encryption [17].

At present, data sets based on homomorphic encryption have been applied to many conventional data mining models and achieved excellent results. Nevertheless, compared to the processing on the plaintext, the calculation time is greatly extended due to the complexity of the model. For example, in Xie's experiments, the neural network after homomorphic encryption required up to an hour to obtain the predictions, while the traditional neural network can complete the prediction in an instant [18].

2.3. Restricted Publication Technology

The privacy protection technology based on restricted publication selectively releases the initial data to achieve the purpose of hiding sensitive information. At present, the commonly used restricted publishing technologies are K-Anonymity [19], L-Diversity [20], and T-Closeness [21], and M-Privacy models [22]. Unfortunately, attackers can use the quasi-identifier attributes and background knowledge in the data set to deduce the privacy information in the restricted data set. Consequently, enhancing the privacy protection intensity of restricted publishing technology has been a hot topic in the security field. In the era of big data, the more prominent research results of improving the technology of restricted publications are presented as follows.

2.3.1. Restricted publication technology based on distributed computing. Restricted publication technology can be roughly divided into two categories: global generalization and local generalization. Global generalization generalizes the quasi-identifier attribute in the data table to be published to the same level [23]. Local generalization generalizes the attributes of the data table to be published to different levels [24]. Local generalization minimizes information loss, which is more flexible than global generalization. With the increasing volume of data, parallel distributed can be used to solve the problem that it is difficult for a single computer to generalize data for its time-consuming part quickly.
MapReduce is a large-scale data processing framework that provides powerful parallel computing capabilities for big data applications [25]. Liu Jie introduced the MapReduce model to the sampling generalization path K-Anonymity [26]. This method combines the advantages of MapReduce and sampling generalization algorithms to reduce the amount of information loss in the published dataset, which also improves the availability of the data. On the other hand, a large number of cloud service applications require users to share their sensitive data, such as electronic medical records, for data mining, bringing the risk of privacy leakage. Zhang proposed a scalable two-stage top-down approach that used the MapReduce framework to anonymize large-scale data sets in the cloud computing environment [27]. In the two phases of the method, Zhang designed a new set of MapReduce jobs to complete parallel computing in a highly scalable manner, effectively improving the scalability, time efficiency and data accuracy of the restricted publication technology.

2.3.2. Restricted publication technology for streaming data. Streaming big data is basically real-time, such as sensor data, call center records, and medical images. Hence, streaming big data processing has higher requirements for latency, accuracy, and real-time processing. A popular method to restrict the publication of static data is K-Anonymity, but K-Anonymity is not suitable for streaming data. Besides, K-Anonymity is an NP-hard type problem. The data sets must be scanned repeatedly during the anonymization process to reduce the loss of information, which is inoperable in stream processing. In particular, the larger the data size, the greater the challenge to the anonymization algorithm.

To address the above challenges, Hessam proposed the FAANST algorithm [28]. FAANST processed data in batches and maintained only K-Anonymity clusters which have been published. Additionally, FAANST had a caching mechanism to publish K-Anonymity clusters, which reach the time limit. However, FAANST is not universal because it supports fewer data types. Guo proposed the FADS algorithm based on the FAANST algorithm [29]. FADS reduced the time spent searching for clusters and improved performance by limiting the number of K-Anonymity clusters that have been released for maintenance. Also, FADS could support the L-Diversity publishing principle by modifying the publishing strategy of tuples, which significantly reduces the information loss.

3. Privacy-Preserving in Image Processing

With the rise of multimedia social networks, some resource-constrained owners will tend to outsource complex image processing processes to cloud service providers. It is worth pondering that if the original image contains sensitive information about the owner, outsourcing the original image directly to an untrusted cloud service provider will lead to the leakage of user privacy. Therefore, privacy-preserving is one of the essential research contents of image security applications. The following is the introduction of significant privacy-preserving research works in image processing.

3.1. Image recognition model based on differential privacy

In the current scientific researches, the researches combine differential privacy and image recognition that are still in the infancy.

Shokri proposed a distributed training method, which injected noise into the "gradients" of parameters to protect privacy in neural networks. Since the magnitude of the injected noise is proportional to the amount at the training time, the method may consume most of the unnecessary privacy budget [30].

To improve this, Abadi proposed Moments Accountant based on the composition theorem, which tracks privacy spending and enforces applicable privacy protections [31]. However, this method introduced noise into the "gradient" of the parameter during each training. If the training samples are large, it will affect the utility of the model. Besides, adding the same amount of noise to all parameters results in a poor utility of the model in the actual scenario. The main reason is that different features and parameters often have different effects on the model output.
In response to the above problems, Phan integrated differential privacy technology and convolutional neural network to design an adaptive Laplace mechanism [32]. This mechanism measured the relationship between input characteristics and output results. It set a reasonable privacy budget to achieve a balance between the security and availability of data in data mining.

3.2. Image generation model based on differential privacy
The two main factors for the great success of deep learning in the field of image recognition are as follows: (1) large data sets with general distribution, (2) considerable computing power with large GPU clusters. With sufficient computing power, how to obtain high quality and a large number of images was the key to improving the ability of deep learning models.

To supply the demand for large-scale samples, Makhzani proposed a generative model [33]. By sketching the data distribution from a small set of training data, the generative model can sample from the distribution and generates more samples. Generative Adversarial Networks (GANs) and their variants combine the complexity of deep neural networks with game theory to generate high-quality "fake" samples that are difficult to distinguish from real samples.

Although GANs have shown impressive performances in modeling the underlying data distribution, it still has the risk of leaking the privacy information of training samples. Since the adversarial training program and the highly sophisticated model of the deep neural network jointly promote the distribution of training samples, the attacker has a great chance to recover the training samples by repeatedly sampling from the distribution.

Focusing on the security of GAN, Xu proposed GANobfuscator, a data publishing framework based on differential privacy, which provides strict privacy protection for the training data of the image generation model [34]. However, GANobfuscator is dependent on the number of training epochs, as it introduces noise into “gradients” of parameters in every training step. In practice, that could potentially affect the model utility, when the number of training epochs needs to be extended to guarantee the model accuracy.

3.3. Image processing based on homomorphic encryption
Yang studied the image processing scheme based on integer vector homomorphic encryption [35]. The research results are suitable for image matrix and image feature vector encryption, and support image processing operations such as weighted inner product and linear transformation. On this basis, Yang has studied the similarity comparison technology of ciphertext vectors to make homomorphic encryption run through the entire process of image processing.

4. Privacy Preserving in natural language processing
In recent years, the development of big global data is still in an active stage. According to statistics and forecasts from the international authority Statista, the global data volume is expected to reach 41ZB in 2019. People's daily lives will be closely related to data. The openness of government and corporate data will play a vital role in the development of society [36]. However, with the openness of data, there are also difficulties. How to strike a balance between data openness and privacy protection is an urgent problem [37]. The published data sets contain a large amount of unstructured data. Since unstructured data has no visible feature identification, it is difficult for conventional detection methods to find the private information contained in it. In most cases, the sensitive information in the data still needs much workforce to detect.

Given the above problems, researchers attend to apply natural language processing technology to detect sensitive information in unstructured text and desensitize text containing sensitive information. Text classification is the most common application in natural language processing. Generally, according to the pre-designed classification model, each text in the original data set can be divided into one or more categories in conformity with the theme and attributes of different texts. Text classification help users find essential information, especially when dealing with large test sets.

Many scholars have conducted various researches on identifying and classifying sensitive data.
Huang proposed the BILSTM-CRF algorithm for sequence tagging and named entity recognition tasks [38]. This model was prevalent in sequence labeling tasks and could significantly improve the accuracy in sequence labeling tasks. In order to improve the recognition rate of the model with more features, Peng used the LSTM-based word segmentation model for word segmentation and combined the output features of the hidden layer of LSTM with the extracted word meaning features to jointly serve as input features of the linear CRF [39]. However, the two training methods combined with sequence labeling tasks not only make the model lack flexibility in adjusting parameters, but also it is difficult to improve the accuracy of word segmentation and named entity at the same time. In terms of text feature extraction, Lu analyzed the impact of individual cases with negative mutual information on classification results and proposed a new mutual information feature selection processing algorithm [40]. Li Qian proposed a new SVM-based network text segmentation framework, and further proposed an SVM-based classification method for network text [41].

In the current research results, the combination of privacy protection and natural language processing are mainly used to detect unstructured text. On the other hand, a sensitive data recognition model for unstructured text is established by combining named entity recognition technology to identify and desensitize the private information contained in the text.

5. Conclusion
With the promotion of digital strategy, the big data industries have become the new engines for economic and social development, while security threats and risks are also increasing [42]. Privacy-preserving data mining methods play an important role in mining and analyzing massive amounts of data. This paper principally introduces and summarizes the latest privacy-preserving technologies, including data distortion technology, data encryption technology and restricted publication technology. Besides, it describes the research trends for data mining methods based on privacy-preserving in the field of image processing and natural language processing.

In this paper, we conclude three research trends of privacy-preserving data mining methods in the following aspects.
1. Privacy Identification Technology
According to the data sources and the needs of users, privacy identification technology is a very important research direction, especially the sensitive information recognition in Chinese.
2. Privacy Computing Technology
It is difficult for a single privacy computing technology to meet the comprehensive privacy protection requirements. Therefore, it is necessary to study multiple privacy computing technologies and combine their advantages to design an integrated privacy computing technology.
3. Privacy Measurement Technology
With the promotion and application of data mining in various industries, there should be more industry-specific privacy measurement technologies to strengthen data security precautions for specific industries and improve the level of vital risk identification and analysis.

References
[1] X. Meng, X. Ci. Big data management: Concepts, techniques and challenges [J]. Journal of Computer Research & Development, 2013.
[2] Colin Tankard. Big data security [J]. Network Security, 2012(7): 5---8.
[3] Elisa Bertino. Big data security and privacy [C]// 2016 IEEE International Conference on Big Data (Big Data). IEEE, 2016.
[4] Agrawal R, Srikant R. Privacy-preserving data mining [C]// Proceedings of the 2000 ACM SIGMOD international conference on Management of data. ACM, 2000.
[5] Yin Y, Kaku I, Tang J. Privacy-preserving Data Mining [J]. Acm Sigmod Record, 2000, 29(2):439-450.
[6] Dasseni E, Verykios V S, Elnagarmid A K. Hiding Association Rules by Using Confidence and Support [J]. Lecture Notes in Computer Science, 2001, 2137:369-383.
[7] Sakuma J. Privacy Protection in Data Analytics [J]. 2014, 6(11):625-635.
[8] Dwork C. Differential Privacy [J]. Lecture Notes in Computer Science, 2006, 26(2): 1-12.
[9] Dwork C. Differential Privacy: A Survey of Results [M]. Theory and Applications of Models of Computation. 2008.
[10] Geng Q, Viswanath P. The Optimal Noise-Adding Mechanism in Differential Privacy [J]. IEEE Transactions on Information Theory, 2016, 62(2): 925-951.
[11] Friedman A, Schuster A. Data mining with differential privacy [C]. Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. 2010.refrigeration, 2012, 33(1): 18-21.
[12] R L Rivest, L Adleman, and M L Dertouzos. On data banks and privacy homomorphisms [M]. Foundations of Secure Computation, Academia Press, 1978.
[13] T. Graepel, K. Lauter, and M. Naehrig. ML confidential: Machine learning on encrypted data [C]//Information Security and Cryptology (ICISC), 2012, pp. 1–21.
[14] Wu, David and Haven, Jacob. Using homomorphic encryption for large scale statistical analysis. 2012.
[15] N. Dowlin, K. Laine, K. Lauter, M. Naehrig, and J. Wernsing. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy [C]// Proceedings of The 33rd International Conference on Machine Learning, 2016, pp. 201–210.
[16] H. Chabanne, A. de Wargny, J. Milgram, C. Morel, and E. Prouff. Privacypreserving classification on deep neural net work. Cryptology ePrint Archive, Report 2017/035, 2017.
[17] Aono, Y., Hayashi, T., and Wang, L. Scalable and secure logistic regression via homomorphic encryption [C]//Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy (2016), ACM, pp. 142–144.
[18] P. Xie, M. Bilenko, T. Finley, R. Gilad-Bachrach, K. Lauter, and M. Naehrig. Crypto-nets: Neural networks over encrypted data arXiv:1412.6181, 2014.
[19] Sweeney L. k-ANONYMITY: [J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2008, 10(05):557-570.
[20] Ashwin Machanavajjhala, Johannes Gehrke. l-Diversity: Privacy Beyond k-Anonymity [C]// Proceedings of the 22nd International Conference on Data Engineering, ICDE 2006, 3-8 April 2006, Atlanta, GA, USA. IEEE, 2006.
[21] Ninghui Li. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity [C]// Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on. IEEE, 2007.
[22] Goryczka S, Xiong L, Fung B C M. m-Privacy for Collaborative Data Publishing [C]// 2013.
[23] EL EMAM K, DANKAR F K, ISSA R. A globally optimal k-anonymity method for the de-identification of health data [J].Journal of the American Medical Informatics Association:JAMIA, 2009, 16(5):670-682.
[24] AMIRI F, YAZDANI N, SHAKERY A. Hierarchical anonymization algorithms against background knowledge attack in data releasing [J].Knowledge-Based Systems, 2016, 101(c):71-89.
[25] MOHAMMED E A, FAR B H. Applications of the MapReduce programming framework to clinical big data analysis:current landscape and future trends [J]. Biodata Mining, 2014, 7(1):1-23.
[26] Quang Tran, Hiroyuki Sato. A Solution For Privacy Protection In MapReduce [C]// Computer Software and Applications Conference (COMPSAC), 2012 IEEE 36th Annual. IEEE, 2012.
[27] Zhang X. A Scalable Two-Phase Top-Down Specialization Approach for Data Anonymization Using MapReduce on Cloud [J]. IEEE Transactions on Parallel and Distributed Systems, 2014, 25(2): 363-373.
[28] H. Zakerzadeh, S. L. Osborn. FAANST: fast anonymizing algorithm for numerical streaming data [M], Springer, 2011, 36-50
[29] K. Guo, Q. Zhang. Fast clustering-based anonymization approaches with time constraints for data streams [J]. Knowledge-Based Systems, 2013, 46:95- -108.,
[30] Shokri and V. Shmatikov. “Privacy-preserving deep learning.” in CCS’15, 2015, pp. 1310–1321.
[31] Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, “Deep learning with differential privacy,” arXiv:1607.00133, 2016.

[32] Phan N H, Wu X, Hu H. Adaptive Laplace Mechanism: Differential Privacy Preservation in Deep Learning [J]. 2017.

[33] Makhzani, Alireza, Shlens. Adversarial Autoencoders [J]. Computer Science, 2015.

[34] Xu C, Ren J, Zhang D. GANobfuscator: Mitigating Information Leakage under GAN via Differential Privacy [J]. IEEE Transactions on Information Forensics and Security, 2019: 1-1.

[35] Xun Yi, Yanchun Zhang. Equally contributory privacy-preserving k-means clustering over vertically partitioned data [J]. Information Systems, 38(1):97---107.

[36] Ning Cao, Cong Wang, Ming Li. Privacy-Preserving Multi-Keyword Ranked Search over Encrypted Cloud Data [C]// INFOCOM 2011. 30th IEEE International Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies, 10-15 April 2011, Shanghai, China. IEEE, 2011.

[37] Cong Wang, Qian Wang, Kui Ren. Privacy-Preserving Public Auditing for Data Storage Security in Cloud Computing [C]// 2010 Proceedings IEEE INFOCOM. IEEE, 2010.

[38] Cynthia Dwork, Kobbi Nissim. Privacy-Preserving Datamining on Vertically Partitioned Databases [C]// Annual International Cryptology Conference. Springer, Berlin, Heidelberg, 2004.

[39] Huang Z, Xu W, Yu K, et al. Bidirectional LSTM-CRP Models for Sequence Tagging [J]. arXiv: Computation and Language, 2015.

[40] Peng N, Dredze M. Multi-task Multi-domain Representation Learning for Sequence Tagging [J]. arXiv: Computation and Language, 2016.

[41] Yu H, Vaidya J, Jiang X. Privacy-preserving svm classification on vertically partitioned data [J]. 2011, 3918:647-656.

[42] Shixian Li, Lei Zhang, Bo Han. A SVM-based compound-word recognition method in information security [C]// 2013 10th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD). IEEE, 2013.