A Survey on the Methods to Determine the Sensitivity of Textual Documents: Solutions and Problems to Solve

Saturnino Job Morales Escobar[1](✉), José Ruiz Shulcloper[2], Cristina Juárez Landín[3], Osvaldo Andrés Pérez García[4], and José Sergio Ruiz Castilla[5]

[1] Centro Universitario UAEM Valle de México, Universidad Autónoma del Estado de México (UAEM), Atizapán de Zaragoza, Estado de México, México

[2] Head of the Research Group on Logical Combinatorial Pattern Recognition, Vice Rectory of Investigations, University of Informatics Sciences, Havana, Cuba
jshulcloper@uci.cu

[3] Centro Universitario UAEM Valle de Chalco, Universidad Autónoma del Estado de México (UAEM), Valle de Chalco Solidaridad, Estado de México, México

[4] Equipo de Investigaciones de Minería de Datos, CENATAV - DATYS, La Habana, Cuba
osvaldo.perez@cenatav.co.cu

[5] Centro Universitario UAEM Texcoco, Universidad Autónoma del Estado de México (UAEM), Texcoco, Estado de México, México
jsruizc@uaemex.mx

Abstract. The identification of sensitive information, whether personal or institutional, is a fundamental step when dealing with the problem of information leakage. This problem is one of the most pressing to which companies and research centers dedicate a considerable amount of material and intellectual resources, as a particular case, to the development of methods or the application of some already known ones to the identification of sensitive information. This increased the proposals with promising results, but without yet offering a totally satisfactory solution to the problem. Under these conditions, it is considered necessary to make a critical analysis of the existing methods and techniques and their future projections. In this paper, a review of the proposals for the determination of sensitivity in textual documents is presented and a taxonomy is introduced to better understand the approaches with which this problem has been approached in the context of information leakage. Starting from the critical analysis and the practical needs raised by experts in the areas of possible application, lines of research on this subject are outlined that include the development of methods for the automation of the classification of sensitive textual documents. Possible extensions that these studies may have in similar application areas are proposed based on other information carriers, such as the cases of images, recordings and other forms of information object, each of which entails levels of complexity that merit studies analogous to the one carried out in this work.

Keywords: Information leakage · Document sensitivity · Information protection systems · Information objects · Supervised classification

Y. Hernández Heredia et al. (Eds.): IWAIPR 2021, LNCS 13055, pp. 280–289, 2021.
https://doi.org/10.1007/978-3-030-89691-1_28

1 Introduction

One of the most valuable resources for any organization is undoubtedly the information contained in its information objects, a concept presented in [1], which correspond to data, documents, images, videos, audio, etc. These information objects, due to the very nature in which they are presented, processed, sent or stored, must be treated in different ways.

Obviously, each information object will possess, either intrinsically or because of the context where it is generated, or in combination of both, a different valuation. It is from this assessment that pertinent actions must be taken for its protection, both to preserve it and prevent its loss, and to prevent its dissemination or access by unauthorized instances.

Textual information objects or documents that can be considered sensitive are of particular interest, but what should be understood by a sensitive document? A document will be sensitive if it contains sensitive information and sensitive information "is that which cannot be made public" [2], or "the sensitivity of the information can be evaluated based on the impact that may result from its leakage" [3]. In the previous definitions it is assumed that once the sensitivity is determined it will not be modified, however, it is common for it to occur. Therefore, for the authors of this work, *the sensitivity of a document is an assessment of its importance, privacy and confidentiality at a given moment*.

Based on the above, given the sensitivity of some information objects, they should be restricted to use, however, due to practical needs, they are used in daily activities, automated or not, making them vulnerable to their theft or inappropriate use.

On the other hand, in the processes of generation, handling or storage of sensitive information objects, there is no certainty that all the organization's personnel follow the security policies and/or that, when using assurance applications, Users comply with the instructions to prevent and avoid unauthorized access. In the case of information leakage, many incidents have been reported in the specialized and dissemination literature. From, for example, the filtering of emails presented in [3], social engineering attacks [4], to the dissemination in the international press and on the internet site of information classified as secret from governments and organizations, negatively impacting them on a social level, economic and political.

According to what is expressed in [5–7], the leakage of information can be the result of deliberate actions or spontaneous errors, which can be increased by its internal or external transmission via email, instant messages, web page forms, among other means and even more, the risk increases when sensitive information objects are shared by customers, business partners, external employees, cloud storage, etc.

In [8] the authors define data leakage as the accidental or inadvertent distribution of sensitive data to an unauthorized entity. Sensitive data for an organization includes intellectual property, financial information, patient information, personal data, among others.

Under these conditions, with the intention of solving this problem, methods have been proposed to ensure data privacy [9], developed systems such as those aimed at Data Leakage Prevention (DLP), detection of data leakage [10], among others, which are designed to detect, monitor and protect confidential data and detect its misuse based on predefined rules. DLP systems are added to traditional security measures such as

Intrusion Detection Systems IDS, which work adequately for well-defined, structured and constant data, as expressed in [5].

Motivated by the relevance, timeliness and complexity of the problem of determining the sensitivity of information objects to face the problem of information leakage, this work presents a critical analysis of the methods for determining the sensitivity of a type of information objects: documents, and we introduce a taxonomy that will help to better understand the approaches with which this problem has been approached and to envision possible lines of work on the subject.

2 Systems for Data Leakage Prevention (DLP)

It is in DLP systems, where most work has been done on the problem of determining data sensitivity and the development of tools that detect and protect sensitive information continues, automatic methods capable of detecting sensitive data and determine the relevant mechanisms to protect them based on their sensitivity are required. Currently, the leakage of sensitive information objects is considered an emerging problem of threat to the security of organizations given that the number of incidents continues to grow.

DLP systems belong to the set of security technologies designed with the purpose of automatically preventing the leakage or loss of sensitive data in any of its three states: in use, in transit or at rest, in the event of problems related to threats [5].

The basic architecture of DLP systems is made up of three modules (see Fig. 1). The first detects whether a document is being sent, created or accessed (for printing, copying, editing, sending over the network, etc.) regardless of its content. The second module analyzes the document detected in the filter, reviews it and sends it to the third module for an assessment in accordance with the established policy. This last module responds by allowing or blocking, if necessary, actions on the information to be protected, issuing the corresponding alert.

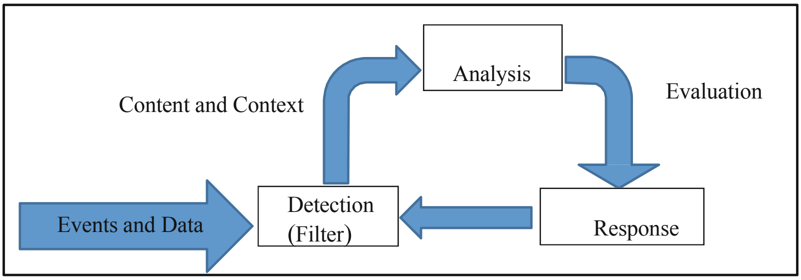

Fig. 1. Basic DLP architecture

The most important characteristics of each module are defined by the policy of the entity that applies the DLP solution to protect its information. For example, the filtering of documents will consider the Threat Model (representation of everything that affects the security of a system) and the Attack Vectors (possible entry gates that can be used for attack) identified entry, information that is domain dependent. The analysis of the detected documents is carried out at the content or context levels, both aspects are closely related to the owner of the information and the files. Finally, the responses that the system

Table 1. Taxonomy of methods used to determine the sensitivity of documents.

Methods and Description
Contextual Approach
Use of metadata associated with sensitive data, e.g., in data submission, source, destination, time, size, format, frequency, topic registration. The metadata can be used in processes or in transaction patterns and is based on defined policies. Examples of proposed algorithms to obtain characteristics that detect misdirected emails can be found in [12, 13] or algorithms based on file sharing between peers [15].
Content Approach
Regular expressions: Set of terms or characters used to form detection patterns, typically used for partial or exact detection in social security numbers, credit cards, personal and corporate records. Specific dictionary-based techniques can speed up and improve detection significantly [14,15].
Classifiers: They depend considerably on an adequate classification of the data. Typically, the owner of the data is responsible for determining the sensitivity of the data and whether it should be protected. Most solutions are based on tag and dirty word list. It is also assumed that to allow access to sensitive data or move it between different domains, all data must be well labeled with its corresponding classification [16-18, 25].
Information object fingerprinting: They are used especially in unstructured data to detect partial or exact coincidence. It is the most common technique used to detect information leakage, DLP with hashing functions such as MD5 and SHA1 can achieve up to 100% accuracy if the files are not altered [19]. Proposals have been made to overcome human or application oversights and maintain detection of sensitive data in transit, using a fuzzy marking algorithm discussed in [20].
N-grams: They are widely used in natural language processing, in machine learning and in information retrieval by weight of terms. It mainly depends on the frequency analysis of terms and n-grams in the documents. The first to use it on DLP were Hart and Johnson to classify business documents into sensitive and non-sensitive; They use Support Vector Machines (SVM) to classify three types of data: private business, public and non-business [15,19, 21].
Weighing of terms: The weighing of terms is a statistical method that indicates the importance of a term in a document, used in text classification and models of vector spaces where documents are treated as vectors and functions are used to determine the frequencies of the terms [22,23].
Machine learning: They use a characteristic space model, where a text sample is transformed into a representation by means of a vector. They use a parameterized function on the training set to make the classification decision. The training sample is divided into groups (clusters) and a classification model based on each of these groups is constructed based on the paragraphs of the document [3, 24].

must issue will be nuanced by the level of security that you want to obtain with the DLP application, which is expressed in the security policy defined for the system.

From the technical point of view, the complexities associated with the Detection and Response modules have been identified [11] and are closely linked to the technological support of the computer system on which the DLP is implemented. It is in the Analysis module where the theoretical problems related to determining the sensitivity of the information to be protected are located.

From the analysis of the DLP systems studied and published methods, the methods can be grouped for study according to the approach assumed by DLP. In Table 1, we present a taxonomy of the methods under these considerations. These methods, for the most part, have been developed with the aim of evaluating the sensitivity level of the information, regardless of when the event that links the document to a security threat occurs. They are oriented to process all the content, assuming that the classification will be Boolean: Yes or No.

3 Analysis of Methods for Determining the Sensitivity of Documents: Advantages and Limitations

Regarding the main context and content methods mentioned, an analysis is presented below with the intention of identifying the advantages and limitations of each of them. In the following section, on this basis, the problems that we consider are pending to be addressed are described.

3.1 Context Analysis

In the development of analysis methods based on the context [12, 15], features such as: file name, file owner and assigned permissions, network protocols, encrypted file formats, user role are used, web services used, web addresses, information associated with the USB type devices used (example: manufacturer, model number) or the desktop application used to edit, read or send the information. With this knowledge, the work of discovering possible channels of information leakage can be guided by applying anomaly detection.

Generally, in this anomaly-based approach, data on the behavior of legitimate users are collected, and then statistical tests are applied to compare it to observed behavior. Based on this comparison, it is determined whether it is legitimate or not. The main element of this approach is the generation of rules that can reduce the proportion of false alarms, both in the detection of new attacks and of already known attacks.

Among the advantages of the approach is the use of features in the description of the information object to determine its sensitivity. But what we consider its greatest disadvantage is that the sensitivity of the document is not determined, it focuses on characterizing the users.

3.2 Content Analysis

Sensitivity based on content is linked to the meaning that the data may have. Each piece of data can contain a large amount of information, however, it can be increased or decreased if it is related to other data. The authors of this work consider that the content of an information object must be closely related to the environment in which said object originates, for example, in a banking institution, a string formed by combinations of 8 characters, is not sensitive, but related to the strings "key" or "access" would change to sensitive data. The following sections present methods used to address this challenge.

Regular Expressions

This method is based on the Theory of Computation where regular expressions are used to represent regular languages, in general terms, the alphabet is identified to form strings, which are used to create detection patterns. These patterns are called regular expressions (RE) and are used by search engines in word processing to validate, generate, extract or replace data. They are typically used for partial or exact detection of social security numbers, credit cards, personal and corporate records. With dictionaries on specific fields, they can speed up and improve the detection of sensitive data significantly.

In [14], RE are used to detect the appearance of critical patterns in the payloads of network packets, for which RE comparisons must be made in real time, and even when Deterministic Finite Automaton (DFA) perform the operation in a linear order time, traversing at most 2N states of the automaton when processing a string of length N, memory requirements can make their use prohibitive despite the improvements presented.

However, applied it in context-based intrusion detection systems have obtained satisfactory results.

In addition to the disadvantage that they comment in [14, 24], on the complexity in space, from our point of view other important limitations are: the complexity of expressing the requirements by means of an RE, they can only be used for regular languages, the application to strings of the language and the difficulty to represent context.

Classifiers

This section reviews some proposals for the evaluation of the sensitivity of information objects, seen as a problem of supervised classification. In [15, 18] the authors emphasize the application of a particular tool and approach, the methods of statistical processing of natural language and the use of machine learning and classification algorithms as decision trees, k -Nearest neighbors, Naive Bayes, Support Vector Machine, text classification algorithms, neural networks, n-grams, among others [25–32].

In general terms, among the limitations observed are: they only consider two classes of documents with a high and low level of confidentiality, or what would be two levels of sensitivity for us; they close the possibility of other kinds of documents; they require classification lists configured manually by the user, which implies that mistakes can be made. On the other hand, most only allow traits of numerical types, they use vectors (vector spaces) for the representation of objects, which excludes qualitative traits and the impossibility of incorporating criteria for the comparison of objects in terms of qualitative traits.

Statistical Analysis

In [19] they present an analysis of information object fingerprint methods and identify two main limitations. The first, that the detection of the leak can be avoided by rewriting the sensitive content and the second, because generally all the content of the document is processed, including non-sensitive parts, false alarms are produced. To remedy the above, they propose an extension of the n-gram matching method called k-jump in ordered-n-grams. As a description of the main idea of the method, it starts with the n-grams method, in which n strings (or portions) of a long sequence of text strings are taken, each selection of n-strings is used to calculate its hash function. In the jump-k method in n-grams it is allowed to skip or ignore up to k-elements of the n-grams. This possibility, when considering relationships between strings that are not adjacent, say the authors, which allows adding contextual information that is not achieved with the n-gram method.

Statistical methods usually consider the most frequently appeared content, while the sensitive content may be ignored if only occupy a small part in the training set.

Finally, in the k-jump proposal in n-ordered-grams, the advantage is offered that you can jump k in the n-chains, but now they are arranged in alphabetical form.

One aspect that we question is the use of a k value which there is no justification for its use, on the other hand, it does not work with incomplete information and assumptions are made on the training matrices.

From our point of view, considering only two classes, sensitive or non-sensitive, limits its application to problems in which different levels of sensitivity are required and even use degrees of sensitivity.

By eliminating articles, prepositions, common phrases, important information about the context in which the information object is generated is eliminated. But the most significant of the disadvantages is that in the representation of the documents the semantics of the terms are being overlooked, which we consider to be a deficiency.

4 Problems to Be Solved

In principle, the problem of determining the sensitivity of documents can be posed as a supervised classification problem, where the description of the information objects is given in terms of features that involve both the context and the content.

Allow quantitative, qualitative features or a mixture of both in the description of objects and use comparison criteria between objects on this basis.

The order in which the word combinations are presented must be considered in determining the sensitivity of documents.

On the other hand, although the determination of the sensitivity of a document has been planted in terms of classes, for example, secret, confidential and unclassified, a flexible way to deal with this problem is by assigning a degree of sensitivity in various levels (documents, paragraphs, sentences, words).

The need to use dynamic training matrices has been detected, that is, they can increase or decrease the number of elements in each class and even that the classes are not balanced and that comparison criteria are considered in which subdescriptions of objects can be compared.

It is important that in the method the procedure for the content analysis of the information object can be made independent of the application context and that it considers that the sensitivity of an information object is temporary, which means that the sensitivity of an object today, it may change at another moment or even cease to be.

With all these considerations, the methods for determining the sensitivity of information objects must be totally flexible and easily incorporated into the classification of documents.

5 Conclusions

In this study, which may not be exhaustive in the analysis of the methods and alternative solutions to the problem of determining the sensitivity of information objects, specifically documents, it has been shown that these solutions partially solve the problem and that there is still a long way to go to offer a tool for the automatic determination of the degree of sensitivity of any information object, particularly documents with unstructured information.

The search must continue from different approaches, the application or development of methods that consider all the factors present in the problem in an integral way, and that undoubtedly increase the complexity of the problem.

A problem to be solved is to develop a tool that is independent of the context, in such a way that said context is generated from a training process in which the context and content are bear in mind. The training would be carried out with sensitive and non-sensitive documents determined by the specialist in the application area. Even the possibility must be allowed that these classes are not disjunct.

The problem of automating the classification of the sensitivity of information objects is broader and more complex than it appears and a detailed study of each of the possible information objects, such as images, recordings and other forms of information object, each of which entails levels of complexity that merit studies analogous to those we have begun on the automation of the classification of sensitive texts.

References

1. Morales, S., Pérez, O., Ruiz, J.: Métodos para la determinación de la sensibilidad de documentos: un estado del arte. Serie Gris, Centro de Aplicaciones de Tecnologías de Avanzada, vol. 036, Habana, Cuba (2016)
2. Berardi, G., Esuli, A., Macdonald, C., Ounis, L., Sebastiani, F.: Semi-automated text classification for sensitivity identification. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, pp. 1711–1714. ACM (2015)
3. Alzhrani, K., Ruddy, E., Chow, C., Boulty, T.: Automated U.S diplomatic cables security classification: topic model pruning vs. classification based on clusters. In: Proceedings of the 2017 IEEE International Symposium on Technologies for Homeland Security (HST), pp. 1–6 (2017)
4. Salahdine, F., Kaabouch, N.: Social engineering attacks: a survey. Future Internet **11**(4), 1–17 (2019)
5. Alneyadi, S., Sithirasenan, E., Muthukkumarasamy, V.: A survey on data leakage prevention systems. J. Netw. Comput. Appl. **62**, 137–152 (2016)
6. Wynne, N., Reed, B.: Magic quadrant for enterprise data loss prevention. Gartner Group Research Note (2016)
7. Ahmad, N.: Do data almost always eventually leak?: Computer **54**(2), 70–74 (2021)
8. Shabtai, A., Yuval, E., Lior, R.: A Survey of Data Leakage Detection and Prevention Solutions. Springer, Boston (2012). https://doi.org/10.1007/978-1-4614-2053-8

9. Jena, M.D., Singhar, S.S., Mohanta, B.K., Ramasubbareddy, S.: Ensuring data privacy using machine learning for responsible data science. In: Satapathy, S.C., Zhang, Y.-D., Bhateja, V., Majhi, R. (eds.) Intelligent Data Engineering and Analytics. AISC, vol. 1177, pp. 507–514. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-5679-1_49
10. Ávila, R., Khoury, R., Khoury, R., Petrillo, F.: Use of security logs for data leak detection: a systematic literature review. Secur. Commun. Netw. **2021**, 1–29 (2021)
11. Wadkar, H., Mishra, A., Dixit, A.: Prevention of information leakages in a web browser by monitoring system calls. In: Proceedings of the 2014 IEEE International Advance Computing Conference (IACC), pp. 199–204 (2014)
12. Liu, T., Pu, Y., Shi, J., Li, Q., Chen, X.: Towards misdirected email detection for preventing information leakage. In: Proceedings of the 2014 IEEE Symposium on Computers and Communication (ISCC), pp. 1–6 (2014)
13. Zilberman, P., Dolev, S., Katz, G., Elovici, Y., Shabtai, A.: Analyzing group communication for preventing data leakage via email. In: Proceedings of 2011 IEEE International Conference on Intelligence and Security Informatics, pp. 37–41 (2011)
14. Becchi, M., Crowley, P.: An improved algorithm to accelerate regular expression evaluation. In: Proceedings of the 2007 ACM/IEEE Symposium on Architecture for Networking and Communications Systems, pp. 145–154 (2007)
15. Sokolova, M., et al.: Personal health information leak prevention in heterogeneous texts. In: Proceedings of the Workshop on Adaptation of Language Resources and Technology to New Domains, pp. 58–69 (2009)
16. Chen, K., Liu, L.: Privacy preserving data classification with rotation perturbation. In: Fifth IEEE International Conference on Data Mining (ICDM 2005), pp. 1–4 (2005)
17. Aggarwal, C.C., Yu, P.S.: A general survey of privacy-preserving data mining models and algorithms. In: Aggarwal, C.C., Yu, P.S. (eds.) Privacy-Preserving Data Mining. ADBS, vol. 34, pp. 11–51. Springer, Boston (2008). https://doi.org/10.1007/978-0-387-70992-5_2
18. Brown, J.D., Charlebois, D.: Security classification using automated learning (SCALE): optimizing statistical natural language processing techniques to assign security labels to unstructured text. Defense Research and Development Canada, Ottawa (Ontario) (2010)
19. Shapira, Y., Shapira, B., Shabtai, A.: Content-based data leakage detection using extended fingerprinting. arXiv prepint arXiv:1302.2028 (2013)
20. Vijayalakshmi, V., Rohini, T., Sujatha, S., Ishali, A.: Survey on detecting leakage of sensitive data. In: World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave), pp. 1–3. IEEE (2016)
21. Hart, M., Manadhata, P., Johnson, R.: Text classification for data loss prevention. In: Fischer-Hübner, S., Hopper, N. (eds.) PETS 2011. LNCS, vol. 6794, pp. 18–37. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-22263-4_2
22. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM **18**, 613–620 (1975)
23. Carvalho, V.R., Balasubramanyan, R., Cohen, W.W.: Information leaks and suggestions: a case study using Mozilla thunderbird. In: CEAS 2009 Sixth Conference on Email and Anti-Spam (2009)
24. Nikitinsky, N., Sokolova, T., Engelstad Ehotskaya, E.: DLP technologies: challenges and future directions. In: The International Conference on Cyber-Crime Investigation and Cyber Security (ICCICS 2014), pp. 31–36 (2014)
25. Engelstad, P., Hammer, H., Yazidi, A., Bai, A.: Advanced classification lists (dirty word lists) for automatic security classification. In: Cyber-Enabled Distributed Computing and Knowledge Discovery, pp. 44–53. IEEE (2015)
26. Kowsari, K., Jafari, M., Heidarysafa, M., Mendu, S., Barnes, L., Brown, D.: Text classification algorithms: a survey. Information **10**(4), 150 (2019)

27. Zorarpacı, E., Özel, S.A.: Privacy preserving classification over differentially private data. Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. **11**(3), e1399 (2021)
28. Guo, Y., Liu, J., Tang, W., Huang, C.: Exsense: Extract sensitive information from unstructured data. Comput. Secur. **102**, 102156 (2021)
29. Patil, D., Lokare, R., Patil, S.: Private data classification using deep learning. In: Proceedings of the 3rd International Conference on Advances in Science & Technology (ICAST) (2020)
30. Trieu, L.Q., Tran, T.N., Tran, M.K., Tran, M.T.: Document sensitivity classification for data leakage prevention with twitter-based document embedding and query expansion. In: 2017 13th International Conference on Computational Intelligence and Security (CIS), pp. 537–542. IEEE (2017)
31. Hassan, F., Sánchez, D., Soria-Comas, J., Domingo-Ferrer, J.: Automatic anonymization of textual documents: detecting sensitive information via word embeddings. In: 2019 18th IEEE International Conference on Trust, Security and Privacy in Computing and Communications/13th IEEE International Conference on Big Data Science and Engineering (TrustCom/BigDataSE), pp. 358–365. IEEE (2019)
32. Lu, Y., Huang, X., Li, D., Zhang, Y.: Collaborative graph-based mechanism for distributed big data leakage prevention. In: 2018 IEEE Global Communications Conference GLOBECOM, pp. 1–7. IEEE(2018)