New Approaches to Data Classification in DLP Systems

Ekaterina Pshehotskaya, Tamara Sokolova and Sergey Ryabov
InfoWatch
Ekaterina Pshehotskaya@infowatch.com, Tamara Sokolova@infowatch.com,
Sergey Ryabov@infowatch.com

ABSTRACT

In this paper, we present possible ways of DLP system development in the future; we also make an introduction to current algorithms and methods widely used in DLP systems and demonstrate our techniques that are supposed to be a solution to some data leakage problems. We describe traffic classification and methods of its analysis depending on data types. We analyze application of specific algorithms and their effectiveness in the aspect of personal data protection. We evaluate our algorithms on our own corpus that we assembled from typical confidential documents; we also demonstrate the difference between the results of standard data processing approach and new data analysis algorithms.

KEYWORDS

Data Leakage Prevention (DLP), Data protection, Information Content Security, Data Mining, Computational linguistics, Machine learning.

1 INTRODUCTION

In today’s world more and more countries adopt laws of personal data protection. DLP systems successfully manage this task [1], [2], but there is an extensive class of personal information, which is poorly or not at all detected by modern DLP systems. Also it is worth noting that there are certain types of textual information that can be detected by usual methods; however there is likely to be a huge amount of false-positive detections. So a system may detect cases that are not a leakage. In this paper we demonstrate some new methods that allow us to detect all data types that may be found in an average company.

In summary, this paper makes the following key contributions to the field of traffic classification:

- We demonstrate our vision of data representation.
- We present our analysis of internal relations between text and binary data. Our algorithm employs new approach to data classification.
- We construct a corpus of typical confidential documents to compare our algorithms with standard DLP analysis algorithms.

The rest of the paper is organized as follows: we describe advantages and disadvantages of standard DLP algorithms in Section 2, and demonstrate what type of data these algorithms can detect. In Section 3 we discuss the DLP software development trends and demonstrate our methods and algorithms; then we describe our test corpus in Section 4. In Section 5 we present the results and, finally, conclude our research in Section 6.

2 DATA LEAKAGE PREVENTION SYSTEMS

In this section, we describe the standard DLP approaches and discuss how they allow the user to prevent the majority of textual information leakage [3], [4], [5]. Let us consider the evolution of traffic analysis tools. Originally DLP systems could detect and analyze only textual data (Figure 1).

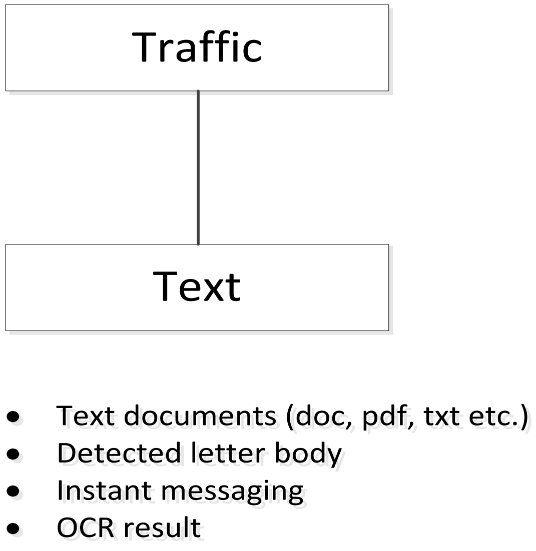

Figure 1

DLP products identified confidential data in three ways: regular expressions, keywords, and hashing [6], [7], [8]. Regular expressions were used primarily to recognize data by type, e.g., social security numbers, telephone numbers, addresses, and other data that had a significant amount of structure. Keyword matching was appropriate when a small number of known keywords could identify private data. For example, medical or financial records might meet this criterion. For less structured data, DLP products used hash fingerprinting. The DLP system took as input a set of confidential documents and computed a database of hashes of substrings of those documents. The system considered a new document to be confidential if it contained a substring with a matching hash in the database. Regular expressions were good for detecting well-structured data, but keyword lists could be difficult to maintain and fingerprint-based methods could miss confidential information if it was reformatted or rephrased for different contexts such as email or social networks [9].

The main disadvantage of this approach was a limited list of protected data: it was impossible to protect, for example, a photo or video recording of any secret document.

Further development of detection and analysis technologies led to binary formatted file detection (Figure 2).

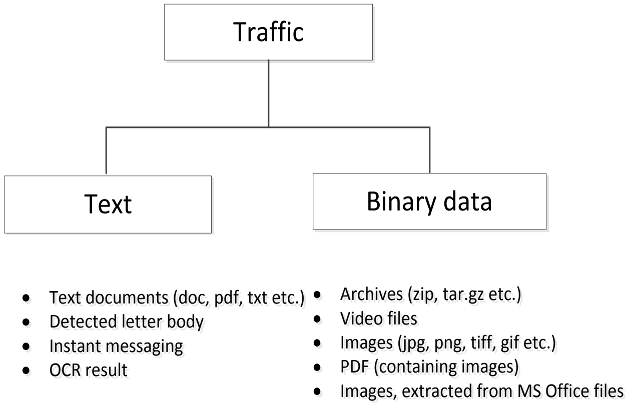
Traffic
Text
Binary data
Text documents (doc, pdf, txt etc.)
Detected letter body
Instant messaging
OCR result
Archives (zip, tar.gz etc.)
Video files
Images (jpg, png, tiff, gif etc.)
PDF (containing images)
Images, extracted from MS Office files

Figure 2

This approach allowed the user to protect a wider range of information, but it was not all-purpose solution because of some characteristic features of binary data. For example, if a user loaded a reference image into the system, and then converted the same image into another format, the system would be unlikely to detect it, due to the internal redeployment of bytes. Thus, this approach would work correctly only for exact matching of analyzed data with reference data. If a user changed any data presentation format, the DLP system would be unable to maintain an adequate level of protection.

Current DLP systems usually single out a new subcategory of binary data – images - to resolve some of the disadvantages described above (Figure 3).

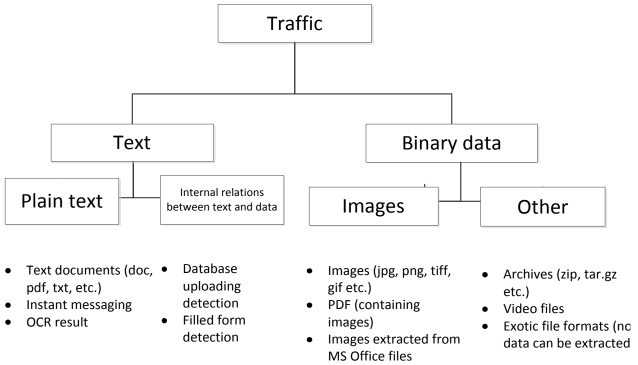
Traffic
Text
Binary data
Plain text
Internal relations between text and data
Images
Other
Text documents (doc, pdf, txt, etc.)
Instant messaging
OCR result
Database uploading detection
Filled form detection
Images (jpg, png, tiff, gif etc.)
PDF (containing images)
Images extracted from MS Office files
Archives (zip, tar.gz etc.)
Video files
Exotic file formats (no
data can be extracted

Figure 3

Thus, current DLP systems use special algorithms that make image analysis format independent; in some cases these algorithms are scale, rotation and noise independent. It is worth noting that some new analysis algorithms are rather demanding of processing power and memory, but modern servers successfully deal with this problem.

Further development in this direction may lead to technologies analyzing new (for DLP systems) file formats. For example, it may be technologies of video content analysis; however, nowadays the capacity of modern processors is not enough for real-time video content analysis.

Another way of analysis technology development is the analysis of internal relations between text and binary data. As an example of such technologies we may name Database Uploading Detection solution and Filled Forms Detection solution.

In order to prevent data leakage via sending photographs or scanned documents the optical character recognition (OCR) is usually used; although it increases the volume of analyzed traffic, it does not prevent the leakage of information that can not be converted into text.

3 CONTENT ANALYSIS

As you may see from the given classifications there are currently several tendencies in the development of content analysis technology:

3.1 Image Analysis

There are several image analysis methods that may be applicable:

3.1.1 Image Classification [13], [14], [10], [12]. This algorithm requires training collection with positive and negative examples. For example, it may be used in order to prevent the leakage of scanned IDs, such as passport, driving license etc.

3.1.2 Sample Stamp Detection [11], [13]. This algorithm assumes that there is a predefined sample stamp; in case any intercepted and analyzed (by DLP system) scanned document contains this stamp, the system will consider it to be confidential and locks it if sent outside the corporate security perimeter.

3.1.3 Image Fingerprints [13]. This algorithm assumes that there is a predefined image sample; in case any intercepted image matches the sample image, the system will prevent this leakage. Image fingerprint algorithm works fine even if the intercepted image has another format or size, or worse quality than sample image.

3.1.4 Scanned or Photographed Credit Card Detection [13], [14], [10].

3.2 Data Internal Relation Detection

This is specialized analysis technique considering data interrelations. Such analysis may be used in order to prevent great number of false positives, which may be obtained as the result of standard DLP analysis methods usage. For example:

3.2.1 Database Uploading Detection. It is impossible to imagine any modern enterprise without large databases. These databases often contain either confidential information or commercial secrets. It would be logical to assume that if we protected the database from leakage, we would minimize data leakage risks. However, such approach would require significant computational resources (databases are usually large, and the amount of confidential information is usually small, so databases contain not so much confidential data), and it would not allow us to protect visual information (some data are not stored directly in a database). There usually is a small limited group of people (e.g. database administrators) having full access to a database while end users can only interact with the database using standard interfaces and receive only certain small amount of data. Considering everything mentioned above we could propose the solution that would help us to protect databases from uploading. The main purpose of the technique is to protect exact data that may leak. However, it is worth noting that some data are not confidential themselves, but they become confidential in conjunction with some other data. For example, the list of employees may not be confidential; however, the list of employees along with their salaries may be considered a commercial secret. Therefore it would be reasonable to introduce a mechanism that could allow us to define relations between the columns of DB. All these actions are likely to improve accuracy and reduce the number of false positives.

3.2.2 Filled Form Detection. In general, the standard digital fingerprint detection algorithm seems to be enough to detect filled forms. However, it detects both filled and empty forms. Therefore, there is likely to be a lot of false positives as empty forms containing no confidential data will be detected. Another problem is that the majority of forms contain such typical input fields as *First Name*, *Second Name*, and *Address* etc. Thus, there is the high possibility of false positives to occur while other non-confidential texts containing these fields may be detected. In order to improve the detection quality of a DLP system it is recommended to use a more complicated solution which would consider relative position of the form fields and allow us to determine whether the intercepted text is a form.

3.3 Multilevel Analysis

Multilevel Analysis may be the best method to solve the problems mentioned above. In order to decide whether intercepted data is confidential, DLP systems usually utilize limited number of techniques, and each analysis result affects the decision separately. In order to improve detection accuracy and reduce number of false positives and false negatives we would recommend you to use Multilevel Analysis. The main feature of this approach is that the decision is reached by combination of results of several algorithms. Multilevel Analysis allows us to flexibly customize the DLP system and use simpler analysis techniques for interception. For example, without Multilevel Analysis it would be possible to protect some types of documents using only classification algorithm, but it would be difficult to configure and might lead to a large amount of false positives and false negatives. Multilevel Analysis allows us to prevent confidential data from leakage using more effective technologies like Filled Form Detection, Database Uploading Detection, Template Analysis (regular expressions) etc. Thus, the final verdict of the system is the mutual result of all analysis algorithms.

4 DLP CORPUS

We have created a corpus for training and evaluating DLP classification algorithms. This corpus contains text documents, images or documents containing both text and images. First, we analysed the corpus using only 3 standard approaches (digital fingerprints, classification and template analysis), and after that we added extended analysis algorithms. The results showed better analysis quality after extended analysis algorithms had been added. For example, in first case (sample form) we had a false positive: empty form was considered a violation. In second case only the filled form was detected (no false positive). On the contrary, analysing DB uploading in the first case we had false negative result (fingerprints are sensitive only to the significant amounts of information). Then, using Database Uploading Detection technique we managed to detect even single string from large database; it would be even possible to detect a confidential part of a string from a database. Then we continued with image analysis: (the corpus contained copies of Russian Federation passports) images were of different quality. OCR method usually fails to extract any text from low quality images; as a result, DLP systems rarely intercept such images. We intercepted about 90% of passports using our image classification algorithm (we trained it on a collection of other passports). The same results were obtained when we analysed credit card scans as well as documents with sample stamps.

5 EVALUATIONS

At the moment we have effective traffic analysis technologies that are likely to prevent the most widespread leakage types that may be encountered in an average organization. Figure 4 displays the classification of algorithms according to the traffic type.

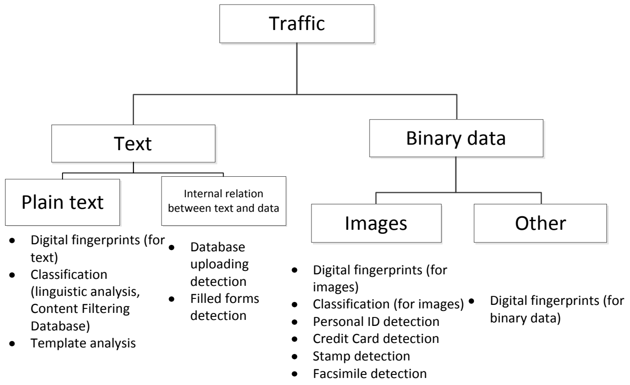
Traffic
Text
Binary data
Plain text
Internal relation between text and data
Images
Other
Digital fingerprints (for text)
Classification (linguistic analysis, Content Filtering Database)
Template analysis
Database uploading detection
Filled forms detection
Digital fingerprints (for images)
Classification (for images)
Personal ID detection
Credit Card detection
Stamp detection
Facsimile detection
Digital fingerprints (for binary data)

Figure 4

Basing on our marked corpus we conducted two experiments. In first one we applied 3 standard DLP techniques (we configured them to intercept maximum possible documents from the corpus – all the required digital fingerprints, text objects and classification categories had been added).

In second experiment we used extended set of DLP techniques (as described above in this paper). All these techniques were also well-configured.

Table 1

Standard		Standard + Extended	
False negatives	False positives	False negatives	False positives
33	4	3	1

In Table 1 you may see the result of these two experiments. As you may note, quantity of false negatives reduced ten times; quantity of false positives reduced as well.

We obtained such results because in second experiment we used techniques analyzing images as images (not as binary files) – and this fact led to higher performance of the DLP system. New methods considering internal relations between different types of data also played a significant role in overall performance improvement.

Thus, usage of the extended techniques set leads to significant performance improvement of a DLP system.

6 CONCLUSIONS

Due to the increasing number of electronic means of sharing information and means of information representation it is necessary to increase the number of methods of confidential data detection. On the one hand, the increasing diversity of data formats handled in analysis technologies leads to the minimization of data leakage risks; on the other hand, special techniques may reduce the number of false positives (this is especially important for high-loaded systems with a large amount of traffic). We see the further development of the DLP industry in these areas. Furthermore, the development of Multilevel Analysis is likely to improve the quality of DLP systems. With help of the mentioned methods it is possible to give a more precise description of any intercepted data type.

7 REFERENCES

[1] T. Hastie and R. Tibshirani. Classification by pairwise coupling. In Proceedings of the 1997 conference on Advances in neural information processing systems 10, NIPS '97, pages 507-513, Cambridge, MA, USA, 1998. MIT Press.

[2] P. J. Hayes and S. P. Weinstein. Construe/tis: A system for content-based indexing of a database of news stories. In IAAI '90: Proceedings of the The Second Conference on Innovative Applications of Artificial Intelligence, pages 49-64. AAAI Press, 1991.

[3] H. Borko and M. Bernick. Automatic document classification. J. ACM, 10(2):151-162, 1963.

[4] F. Sebastiani. Machine learning in automated text categorization. ACM Comput. Surv., 34(1):1-47, 2002.

[5] L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Wadsworth, 1984.

[6] Symantec. http://www.symantec.com/data-loss-prevention.

[7] Websense. http://www.websense.com/content/data-security-suite-features.aspx

[8] McAfee. http://www.mcafee.com/ru/products/total-protection-for-data-loss-prevention.aspx

[9] M. Hart, P. Manadhata, and R. Johnson. Text Classification for Data Loss Prevention, HP Laboratories. HPL-2011 -114.

[10] H. Bay, T. Tuytelaars, and L. Van Gool, “SURF: Speeded Up Robust Features”, 9th European Conference on Computer Vision, 2006.

[11] P. Viola and M. J. Jones. Rapid Object Detection using a Boosted Cascade of Simple Features. IEEE CVPR, 2001.

[12] G. Csurka, C. R. Dance, L. Fan, J. Willamowski and C. Bray, “Visual Categorization with Bags of Keypoints”, 2004.

[13] R.Szeliski, "Computer Vision: Algorithms and Applications", 2010, Texts in Computer Science.

[14] D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints”, International Journal of Computer Vision, 60, 2, pp. 91-110, 2004.