Chapter 7
The Algorithm of Semantic Analysis in Disruptive Information Security Systems

Polina Kisarina
EGAR Technology, Russia

Andrey Mishin
https://orcid.org/0000-0002-4849-0710
Vladimir State University, Russia

ABSTRACT

This chapter represents some of the main drawbacks of DLP systems implemented by businesses in international practice. The main structural shortcomings of these systems have been analyzed, and the factors correlating with them were revealed. An experimental setup has been formed to assess the impact of changes in these factors on the Type 1 and 2 errors in the operation of the systems. The authors also provide the results of the research with the use of algorithms, including the influence of the identified factors in the business systems of different directions to improve the economic security of the company.

INTRODUCTION

A formalized approach to business allows you to clearly structure organizational actions in the form of documentation, some of which involves the display of potential actions of the organization. On the basis of corporate data sets, it is possible to form a set of indicators, which are then converted into models for evaluating the effectiveness of various kinds, and are used as well as insider audit information.

To counter such information security threats many organizations are implementing DLP-systems that allow analyzing data sets of corporate communication. One of the analysis sections in such systems is semantic-syntactic data analysis.

In the analysis of previous studies of such systems (DLP Technology, 1998), (Hart et al., n.d.) there is a tendency – the implementation of such systems can be used as a means of protecting information

DOI: 10.4018/978-1-7998-0361-4.ch007

and preventing its leakage, but the operation of such systems is imperfect, primarily because the work of the automated analyzer is formalized and can not include data analysis in the context, as a result of the operation of such a system there are errors of the first and second kind (namely, errors of the first kind - the missing of confidential information leakage, as well as the high probability of false positives – errors of the second kind).

Based on the above studies, it is obvious that the work of the semantic analyzer within the DLP-system needs some categorization and indexing of data added to it with its correction to the context, which would reduce the errors of false positives of the system, as well as the number of errors of the first kind.

MOTIVATION

In the analysis of previous studies of such systems (DLP Technology, 1998), (Hart et al., n.d.) there is a tendency – the implementation of systems can be used as a means of protecting information and preventing its leaks, but the work of such systems is imperfect, primarily because the work of the automated analyzer is formalized and can not include context data analysis, as a result there are errors of the first and second kind (namely, errors of the first kind - the omission of confidential information leakage, as well as a high probability of false positives – errors of the second kind) in the work of such systems.

Based on the above studies, it becomes obvious that the work of the semantic analyzer within the DLP-system needs additional categorization and indexing of data with their correction to the context, which would reduce the errors of false positives of the system, as well as the number of errors of the first kind. Thus, the reason that initiated this study is an attempt to adapt the work of the final technological products that have already found a commercial form to the specifics of the final organizations with an increase in the quality of processing of data sets.

The ultimate goal of the study can be formulated as an increase in the efficiency of the analysis of official correspondence of the organization in the evaluation of information security.

CONTRIBUTION

In the framework of this study, we propose an experimental study of the algorithm, with the following improvements:

· the possibility of correction of data indexing depending on its location in the text;
· the categorization of the studied data sets by structured dictionaries with a specific signature database

The formulation of the scientific hypothesis is as follows: suppose that a given frequency of occurrence of tokens from signature bases prepared by the organization will more qualitatively indicate a certain level of its information security depending on the categorization of signature bases and correction for its location in the text.

The problem, the solution of which is proposed to be obtained in the work, is defined as optimization, namely: the need to maximize the number of fragments found, that informatively for the researcher cross the perimeter of the given signature constraints while minimizing the errors of false positives.