An approach combining general and highly specialized semantic analysis in DLP systems

1st Dmitry Purtov
Volga State Technological University
Yoshkar-Ola, Russia
idmitry.purtov@gmail.com

2nd Irina Sidorkina
Volga State Technological University
Yoshkar-Ola, Russia
igs592000@mail.ru

***Abstract*—The paper proposes an approach to combining general and highly specialized semantic analysis. The analysis of the main problems of the semantic analysis of the text is carried out, as well as an approach is proposed in which the semantic analysis in DLP systems immediately spreads to the entire protected system. The presented approach will allow to gradually accelerate the work of the DLP system, as well as analyze the result of the work of semantic analysis to evaluate and maintain semantic analysis in an up-to-date state**

***Keywords*—semantic analysis, DLP system, morphological reference book, visualization of semantic analysis.**

The development of modern computer technology has allowed the creation of systems for the analysis of not only structured data, but also data presented in natural language. Typically, automatic text analysis systems use the following analysis steps: morphological, syntactic, and semantic. There are many different tasks where automatic text analysis is needed, such as in DLP systems.

DLP (Data Loss Prevention) — It is a system designed to prevent confidential information from leaking outside the corporate network. This system is built on the analysis of data streams that go beyond the corporate network. In the event of a certain event characterizing the transfer of confidential information, the system either blocks such a transfer or sends notifications to the operator. A typical operation scheme of DLP-systems is presented in Figure 1.

- Means of intercepting information transmitted via external channels (outside the protected automated system). This category includes drivers for controlling the printing of information, drivers for controlling connected devices, firewalls that control network traffic, etc. [1].
- The categorizer that makes up the core of the DLP system. His work is to analyze the transmitted information, as a result of which the category is uniquely determined (degree of information confidentiality) [1].
- Means of response and registration. Based on the degree of confidentiality determined by the categorizer, the DLP system responds in accordance with the system settings and blocks the transfer of confidential information, or the security administrator is alerted (signaling) about unauthorized transmission (leakage) of information [1].

In such systems, semantic analysis is used in the stream analysis of text data and is part of the categorizer. Semantic analysis helps to determine whether there is confidential information in the text, relying directly on the contents of the text, its meaning, and not on specialized labels (vultures) that may or may not be deleted at all. The meaning of the text refers to the text from the point of view of a person, that is, obtaining facts that are clearly present in the text and revealing the hidden meaning from the text if it is present [2].

Semantic analysis is closely related to structural analysis. Both there and there are dependencies of words, analyzes the connection, their strength. But if structural analysis takes into account only the language rules for constructing sentences, word dependencies through parts of speech, then semantic analysis takes into account the meaning of words to all this. Without semantic analysis, an analysis of the text can make a mistake, since the structural representation of the text can have several representations and, in addition, may not convey meaning even if it is correctly constructed from the point of view of language rules. It is worth noting the fact that since the main task of semantic analysis is understanding the meaning, we can conclude that it cannot exist without the morphology of the language. Morphology here refers to various morphological representations of the word, the so-called word forms. To store word forms, various reference books are used. Reference books store signs and words by which the context is identified with subsequent understanding of the meaning of the text.

In semantic analysis, many features arise that cause problems in the implementation of semantic analysis. Here are some problems that need to be addressed in semantic analysis:

Knowledge of context. Words, or rather their meanings can vary greatly depending on the context. The same words, sentences can have completely different meanings, depending on the context of their use.

Various sentence structures. If we consider the natural language in a broad sense, we can see that Latin, Cyrillic

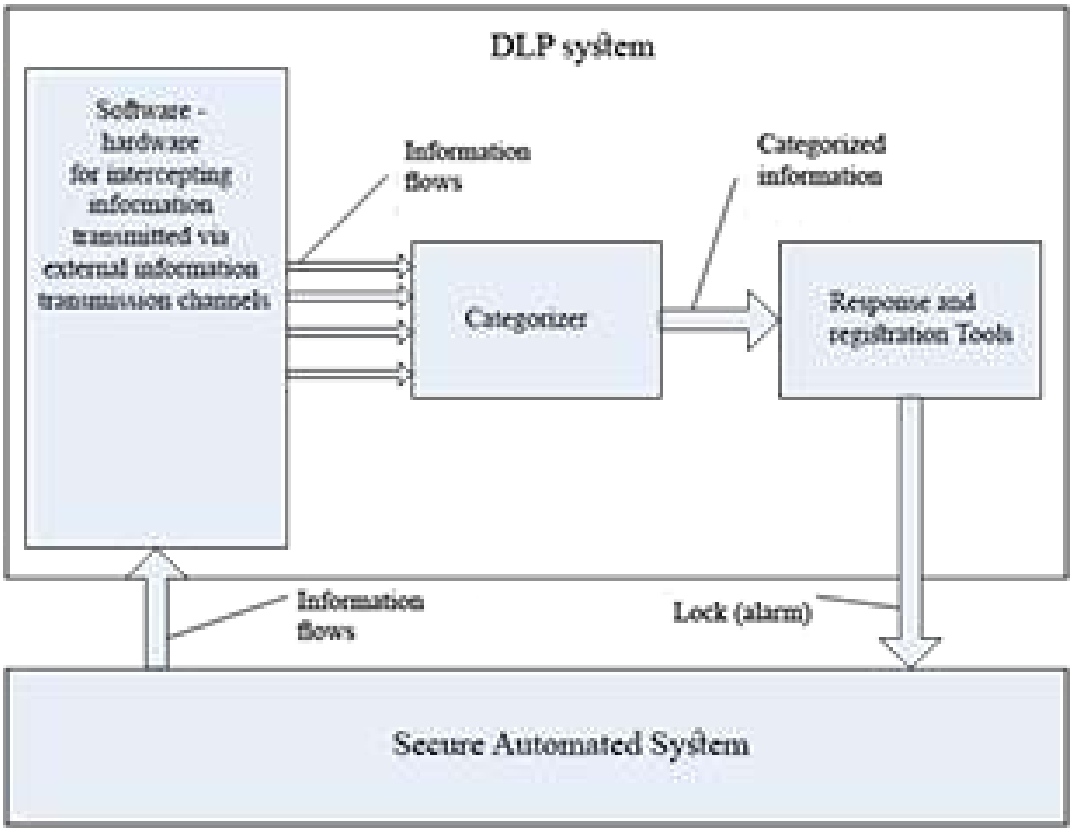

Figure 1. DLP system operation diagram.

languages have their own structure. Data fact does not allow relies on the structure to understand the meaning of the proposal. Yes, and the same sentences can be rewritten so that their structures are different, but the meanings have not changed.

Interpretation problem. One and the same subject can be described in completely different words. In this case, even when using other words, the general meaning has not changed.

The emergence of new words. Over time, new words appear in circulation that can describe existing things or completely new ones.

Contradictions. In a number of languages, it is possible to use seemingly unrelated words to convey a certain meaning. Basically, this phenomenon is observed in works of art when compared, etc.

Ellipses. These are sentences in which words are omitted as their presence is implied in relation to the context.

Semantic analysis has many problems that are quite difficult to solve. Suppose that there are already some solutions to problems, but all of them are not ideal, and there will not be an ideal solution. Many decisions are based on formal grammars. The main representatives of this approach can be called Melchuk I., Chomsky N. and others [3]. At the heart of formal grammar is an attempt to create a new universal language based on mathematical representation [3].

When using semantic analysis in DLP systems, the speed and quality of analysis, or rather the balance between them, is very important, since the analysis takes place in a stream form and these two criteria are mutually exclusive. If we consider semantic analysis in general, we can see that there are systems that use semantic analysis to understand any text in a natural language, such systems are more universal, but slow, as well as highly specialized analysis, which is quite fast, but effective in only one context . DLP systems usually use large digging, where there are many subdivisions, departments, etc. Each of these links usually works strictly with a certain type of data. For example, the economic department with financial documents, and the legal with legal, etc. To increase the speed of analysis of documents from such units, it is more efficient to use highly specialized analysis since the specifics and the context of the information that needs to be checked are already known. But do not forget about units where there is no clear context of information. They can get any kind of information. In such cases, a combination of analysis methods is better. Combination gives a better result, although it complicates the whole system than choosing one analysis method. When combining, the

main thing is to find a balance between quality and speed, correctly choose the methods that analyze the text.

When deploying a DLP system with semantic analysis of texts, the following method can be proposed, which can provide maximum efficiency with minimal delay, as well as reduce the costs of its operation. As an example, take a medical organization, where there are many subdivisions, departments. For all units at the beginning, it is necessary to use universal semantic analysis. One of the representatives of universal analysis can be called UNL (Universal Networking Language). UNL is an artificial language designed to store data [4]. UNL in this case is a kind of intermediary, a universal way of storing data, not tied to any context, that is, when analyzing the text, the text is translated into the UNL representation and then it is analyzed. Upon completion of the analysis, the result, also presented in the UNL representation, is translated into natural language. The main problem of a universal system when using semantic analysis is the need to store a reference based on which the system could identify the context, the meaning of words, etc. In view of the fact that this approach uses large directories and many text transformations, the speed of such an analysis will be very slow. But such a solution will secure the company's infrastructure at the initial stage. To speed up the analysis of the text, it is worth introducing a highly specialized analysis in those units where it is possible, such as extracting information about injuries in medicine [5]. But here it is worth noting one feature, namely, in general, the essence of a highly specialized analysis is almost the same; only the context differs. The context here refers to the area of use of words, rules for building texts, that is, the input and output data vary depending on the unit. Thus, it turns out that a choice arises between a long but universal analysis and a fast but highly specialized analysis. Although there is now a choice between the two approaches, it is still worth using both in parallel.

Let us consider two situations when the text passes through a DLP system, and the DLP system in one case has highly specialized analysis and is universal, and in the other cases there is no highly specialized analysis. When there is no highly specialized analysis, more time is spent on analysis, but at the same time we take into account all options for leakage of confidential information. But when we have a highly specialized analysis, we can save and analyze quickly with a small probability that we will miss the case of information leakage if the text contains information that is not in the context of a highly specialized analysis. Therefore, to reduce risks, it is worth additionally sending a text for a general analysis, but if a highly specialized analysis has passed, you should not wait for the result of a general analysis. In this case, a general analysis is only an additional, indirect check, the result of which we can receive belatedly. In addition, do not forget about the maintenance of the system. Directories need to be updated over time, and incidents should be analyzed for their correctness. To do this, it is worth visualizing the result of the analysis. This is necessary for the timely updating of directories, which are used both in general analysis and in highly specialized ones. Here, visualization is understood as a conclusion to the screen for a detailed analysis of word relationships in the text with the possibility of quick updating of the reference [6]. The main task of visualization is to help and partial automation of the data update process. This will allow, with minimal human resources, updating directories even in role-time mode. For the assessment, it is worth using the number of leaks detected. Moreover, you need to consider the total number, missed and false information leaks. This will help identify bottlenecks in the analysis for further improvement by changing the algorithm or updating directories.

The proposed approach with combined general and highly specialized semantic analysis will secure the corporate network. An increase in security will occur due to streaming semantic analysis of all textual information. This completely eliminates the direct leak of textual information. Thanks to the combined method of text analysis, the speed of information processing increases. The increase in speed can be measured, compared with the difference in the system before the introduction of highly specialized analysis and after. The difference in speed will be different. In addition, the results of the analysis will allow you to quickly update the directories. In reference books there are words that identify the context and meaning, as, for example, in our case, the reference may contain medical illnesses. Handbooks of this kind should be updated with the advent of new diseases. Visualization of the analysis in such cases will allow This is necessary in order to get rid of the need to search for the necessary reference. This takes time so that the administrator can find specialized software.

REFERENCES

[1] Kumunzhiev K.V., Zverev I.N. Method for increasing the efficiency of the dlp-system in the semantic analysis and categorization of information. Modern problems of science and education, 2014, No 5.

[2] Pospelov D. A. Ten hot spots in research on artificial intelligence. Intelligent Systems (Moscow State University), 1996, Vol. 1, No 1-4, pp. 47–56.

[3] Batura T.V. Methods and systems of semantic analysis of texts. Software products, systems and algorithms, 2016, No 4, pp. 45-57.

[4] What is UNL? L. Kreidlin, 2001. Available at: https://old.computerra.ru/2001/390/197284/.

[5] Golovko A.P. Automatic recognition of the meaning of the text on the example of reports on industrial accidents. VESTNIK OF Kostroma State University, 2016, No 3, pp. 106–112.

[6] Purtov D. N. Sidorkina D. N. The problem of training a neural network when extracting key information. Intelligent systems and information technologies, 2019.

Подход с комбинированием общего и узкоспециализированного семантического анализа в DLP-системах

Пуртов Д. Н., Сидоркина И. Г.

Развитие современных компьютерных технологий позволило создавать системы для анализа не только структурированных данных, но также данных, представленных на естественном языке. Как правило системы автоматического анализа текста используют следующие этапы анализа: морфологический, синтаксический и семантический. Подобного рода системы использоваться в DLP-системах. DLP (Data Loss Prevention) -— это система, созданная для предотвращения утечек конфиденциальной информации за пределы корпоративной сети. Наиболее интересной частью анализа текста является семантический анализ. Именно при семантическом анализе выясняется смысл текста. В данной работе были рассмотрены основные проблеме семантического анализа, а также предложен способ построения семантического анализа, который может обеспечить максимальную эффективность при минимальных временных затратах и уменьшит издержки при эксплуатации. Особенность подхода заключается в особом комбинировании общего семантического анализа с специализированным, постепенном увеличении скорости анализа в DLP-системе, визуализации анализа с возможностью улучшения работы семантического анализа. Под общим анализом тут понимается возможность анализировать текст любого рода, а специализированный тексты только определенной тематики. Постепенный плавный переход от общего к специализированному анализу позволяет увеличить скорость работы системы в тех тематических областях. Визуализация дает возможность более быстро анализировать и оценивать результат работы семантического анализа, позволяет сразу актуализировать справочники слов, необходимых для понимания смысла текста. Все эти действия позволяют DLP-системе быть в актуальном состоянии для предотвращения утечки информации.

Received 15.01.2020