УДК 004.056.5

МЕТОД ПОВЫШЕНИЯ ЭФФЕКТИВНОСТИ DLP-СИСТЕМЫ ПРИ СЕМАНТИЧЕСКОМ АНАЛИЗЕ И КАТЕГОРИЗАЦИИ ИНФОРМАЦИИ

Кумунжиев К.В., Зверев И.Н.

Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Ульяновский государственный университет», Ульяновск, Россия (432017, Ульяновск, улица Льва Толстого, 42), e-mail: inz_2008@mail.ru

Предлагается метод повышения эффективности систем защиты информации от утечек (DLP-систем), основанный на более полном использовании семантической информации, имеющейся в тексте. В статье анализируются принцип действия и типовая структура имеющихся DLP-систем. Анализ показал, что наиболее перспективное направление повышения эффективности таких систем – использование семантической информации, имеющейся в тексте. Предлагается двухступенчатая обработка информации. На первом шаге производится категоризация на основе тезауруса текста. В случае если анализ текста с применением тезаурусов дает отрицательный результат, то дальнейший анализ не производится. На следующем шаге, в процессе работы генератор создает онтологию анализируемого текста и передает ее анализатору, который производит процедуру сравнения с онтологией предметной области. Предлагаемый метод позволяет существенно снизить размерность задачи (а, соответственно, и ее трудоемкость) и создавать на его основе DLP-систему, эффективно использующую семантику текста для поиска в нем защищаемой информации.

Ключевые слова: защита информации, DLP, защита информации от утечек, онтологии, тезаурусы.

THE METHOD FOR INCREASING THE EFFECTIVENESS OF DLP-SYSTEM WITH SEMANTIC ANALYSIS AND CATEGORIZATION OF INFORMATION

Kumunzhiyev K.V., Zverev I.N.

Ulyanovsk State University, Ulyanovsk, Russia (4132017, Ulyanovsk, Leo Tolstoy Street, 42), e-mail: inz_2008@mail.ru

Propose a method for enhancing the effectiveness of the protection of information leakage (DLP-systems) based on a more complete use of the semantic information available in the text. The paper analyzes the principle and the typical structure of DLP-systems available. The analysis showed that the most promising way to improve the effectiveness of such systems – the use of semantic information available in the text. It is proposed a two-stage processing of information. In the first step the categorization on the basis of the text of the thesaurus. If text analysis using thesauruses gives a negative result, then a further analysis is performed. In the next step, in the process of ontology generator creates the analyzed text and passes it to the parser, which produces the comparison with the domain ontology. The proposed method can significantly reduce the dimension of the problem (and, respectively, and its complexity), and create on its basis DLP-system effectively uses the semantics of the text to find it protected information.

Keywords: information protection, DLP, data leak prevention, ontology, thesaurus.

В последние годы осложнилась ситуация с внутренними угрозами, в частности, с инсайдерами (злоумышленниками, являющимися членами организации-владельца конфиденциальной информации). В связи с этим для защиты информации от утечек возникло новое направление в области информационной безопасности – так называемые DLP-системы [1].

Интенсивное развитие информационных систем приводит к резкому увеличению внутренних угроз, что, в свою очередь, делает необходимым постоянное развитие DLP-систем.

1. Принцип функционирования DLP-систем

Типовая схема функционирования современных DLP-систем представлена на рис.1.

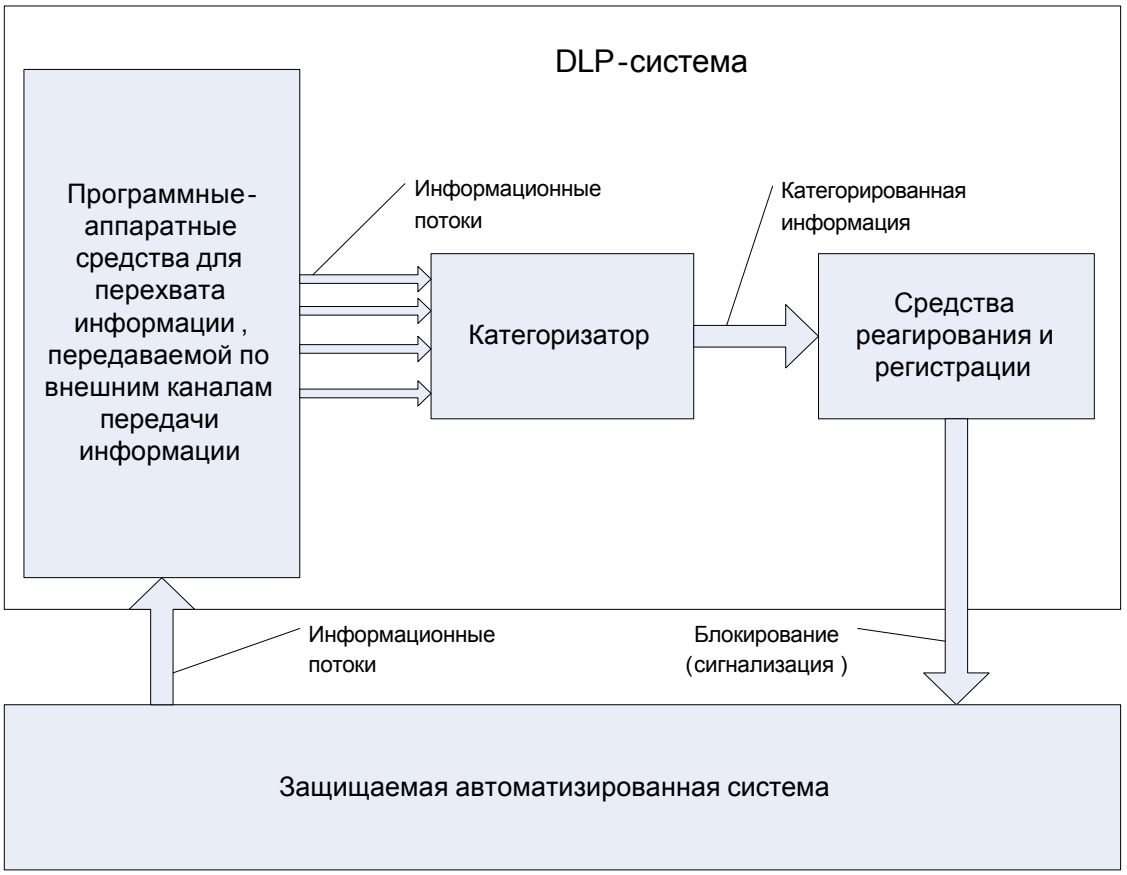

Рис.1. Схема функционирования DLP-системы

Можно выделить 3 основных подсистемы DLP-систем:

1) **Средства перехвата информации**, передаваемой по внешним каналам (за пределы защищаемой автоматизированной системы). К данной категории относятся драйверы для контроля вывода информации на печать, драйвера для контроля подключаемых устройств, межсетевые экраны, контролирующие сетевой трафик и т.д.

2) **Категоризатор**, составляющий ядро DLP-системы. Его работа заключается в анализе передаваемой информации, в результате которого однозначно определяется категория (степень конфиденциальности информации). Процесс определения категории и конфиденциальности информации на основе смысловой близости принято называть **категоризацией информации** [2].

3) **Средства реагирования и регистрации**. На основании определенной категоризатором степени конфиденциальности DLP-система реагирует в соответствии с системными настройками – производится блокирование передачи конфиденциальной информации, либо производится оповещение (сигнализация) администратора безопасности о

несанкционированной передаче (утечке) информации.

Типовая схема работы категоризатора DLP-системы представлена на рис. 2.

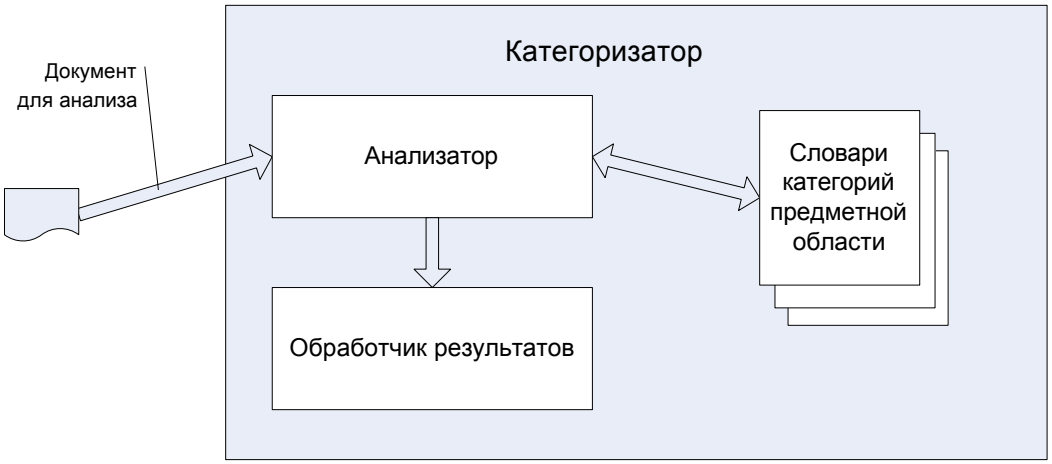

Рис. 2. Типовая схема категоризатора

Выделим основные элементы категоризатора:

1) **Словари категорий предметной области** предназначены для категорирования информации по нахождению в анализируемом тексте определенного количества слов из словаря определенной категории. При создании словарей необходимо привлечение специалистов по лингвистическому анализу.

2) **Анализатор** – ключевой элемент категоризатора, проводящий поиск в анализируемом тексте ключевых слов по словарям для определения принадлежности к той или иной категории. Современные DLP-системы предлагают для повышения эффективности поиска ввод весовых коэффициентов для отдельных ключевых слов в словаре. При анализе используются статистические и вероятностные методы. Как правило, используется метод Байеса для подсчета вероятности того, что анализируемый текст относится к определенной категории [2]. Чаще всего анализ сводится, в лучшем случае, к поиску в тексте ключевых слов по тематическому словарю и категорированию с помощью байесовского метода по заранее установленным весовым коэффициентам.

3) **Обработчик результатов** предназначен для обработки и вывода результатов работы категоризатора информации. Выделение данного элемента обусловлено тем, что возможна ситуация, когда анализируемый текст будет отнесен сразу к нескольким категориям. В этом случае обработчик результатов и должен обеспечить гибкую логику работы в зависимости от пользовательских настроек.

2. Оценка эффективности DLP-систем

Оценку эффективности будем производить по следующим критериям:

1) количество ложных срабатываний или ложных тревог (ошибки первого рода) [5];

2) пропущенные (необнаруженные) утечки информации (ошибки второго рода);

3) трудоемкость (быстродействие) DLP-системы.

Высокое количество ложных срабатываний – главная проблема описанной выше схемы. Это связано со сложностями при работе с отдельными словами естественного языка. Зачастую отдельные слова могут затрагивать совершенно разные категории информации. Чаще всего администратору DLP-систем приходится вручную разбирать огромное количество информации, по ошибке попавшей в защищаемые информационные категории. Метод Байеса, часто применяемый в данной схеме, также приводит к ложным срабатываниям. При его применении исходят из независимости появления слов в тексте, что в корне неверно.

Количество необнаруженных утечек для данной схемы теоретически должно быть небольшим, но при некорректной настройке словарей категорий оно может резко возрасти (например, администратор DLP-системы может «облегчить» себе работу и уменьшить количество ложных срабатываний, удалив значительное количество слов из словаря категории).

Таким образом, при приемлемой трудоемкости описанная выше типовая DLP-система дает недопустимый высокий процент ошибок и в целом является (оказывается) **неэффективной**. Основная причина этого – высокая размерность задачи категорирования.

Анализ показал, что наиболее перспективное направление повышения эффективности таких систем – использование семантической информации, имеющейся в тексте. Предлагается двухступенчатая обработка информации. На первом шаге производится категоризация на основе тезауруса текста. В случае если анализ текста с применением тезаурусов дает отрицательный результат, то дальнейший анализ не производится.

На следующем шаге, в процессе работы генератор создает онтологию анализируемого текста и передает ее анализатору, который производит процедуру сравнения с онтологией предметной области.

Предлагаемый метод позволяет существенно снизить размерность задачи (а, соответственно, и ее трудоемкость) и создавать на его основе DLP-системы, эффективно использующие семантику текста для поиска в нем защищаемой информации.

Далее будет приведено подробное описание предлагаемого метода.

3. Метод повышения эффективности DLP-систем

По сравнению с известными методами использование онтологий дает следующие преимущества:

1) масштабируемость – количество документов в базе защищаемой информации не существенно влияет на время работы алгоритма;

2) сокращение «концептуального несоответствия», т.к. онтология является инструментом, работающим приближенно к человеческому способу мышления;

3) упрощение повторного использования знаний – использование уже определенных в других онтологиях понятий, соответственно возможно использовать уже существующие для данной предметной области онтологий;

4) «семантическая эффективность» – при сравнении семантики текста и онтологии предметной области возможно добиться наибольшей точности в категорировании и существенно снизить количество ложных срабатываний;

5) готовый набор средств для создания (использования существующих) онтологий и задания правил анализа – существующие редакторы онтологий (например, Protege) предоставляют удобный инструментарий для создания и редактирования онтологий предметной области.

Основным недостатком применения онтологий является значительная ресурсоемкость (трудоемкость системы).

Для снижения трудоемкости целесообразно использовать категоризацию по словарям (тезаурусам) смысловых категорий, предложенную и описанную ниже.

Рассмотрим предлагаемую схему онтологического категоризатора, представленную на рис. 3.

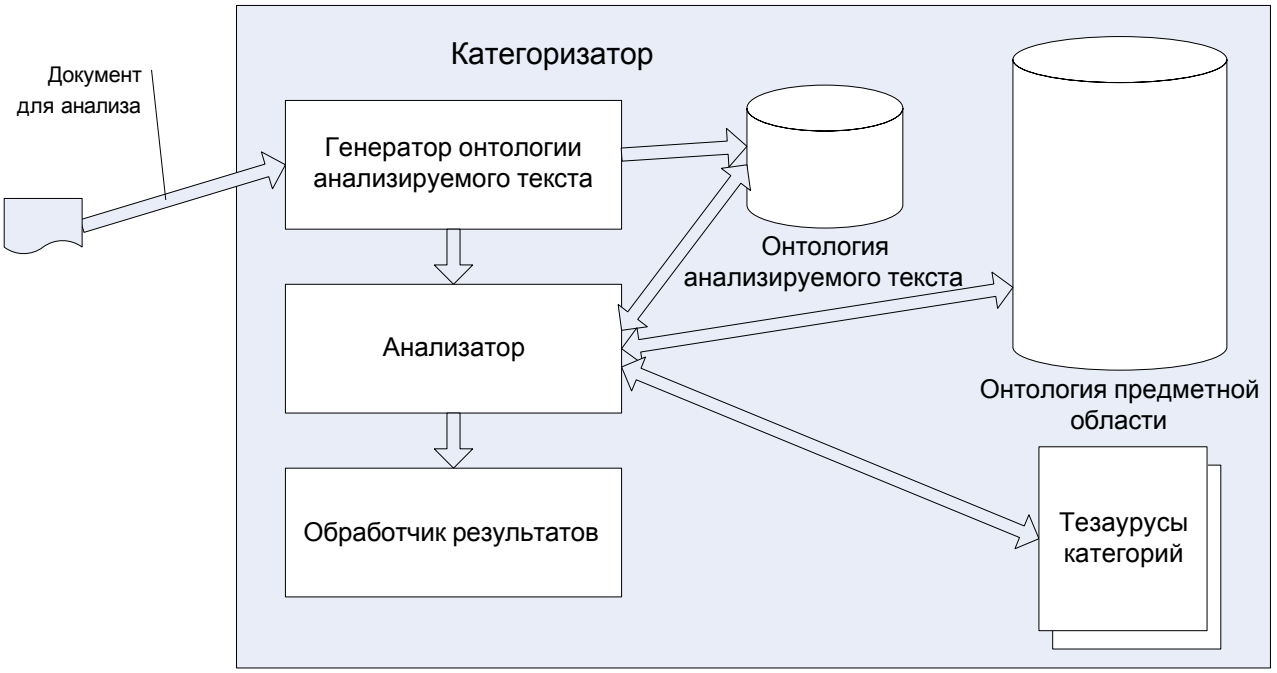

Рис. 3. Схема онтологического категоризатора

Выделим основные элементы категоризатора:

1) **Генератор онтологии анализируемого текста (документа)**. Генератор в автоматическом режиме разбирает анализируемый текст и строит онтологию, отражающую семантику текста.

Заметим, что построение онтологии должно вестись по тем же правилам, что и построение онтологии предметной области;

2) **Анализатор** – проводит сравнение между онтологией анализируемого текста и онтологией предметной областью для определения принадлежности той или иной категории. Для ускорения анализа используются тезаурусы категорий, определяющие, к какой категории точно не относится анализируемый текст. Подробно алгоритм работы анализатора будет описан ниже.

3) **Обработчик результатов** предназначен для обработки и вывода результатов работы онтологического категоризатора.

4) **Онтология анализируемого текста** создается для дальнейшего семантического анализа текста.

5) **Онтология предметной области** создается до ввода в эксплуатацию DLP-системы. В ее создании принимают участие специалисты предприятия, где будет внедряться DLP-система – эксперты по информационной безопасности и специалисты по защищаемым предметным областям. Для создания онтологии используется тот же алгоритм, который применяется для автоматической генерации онтологии анализируемого текста. Разница в том, что создание онтологии предметной области – более сложный и трудоемкий процесс, требующий ручного ввода и участия группы экспертов. Правила, заданные в процессе создания онтологии предметной области, будут применяться затем в автоматической генерации. Наборы правил должны однозначно определять условия отношения к той или иной категории. Онтологии, как правило, определяются триплетами [3]:

[A, p, B] , где A – субъект, p – предикат, B- объект.

В качестве триплета можно привести пример. В анализируемом тексте содержатся сведения о некоем изделии С456, в частности приводится его состав. В процессе автоматической генерации онтологии текста появляются триплеты вида [«Изделие С456», contains, «часть X»]. Если в онтологии предметной области будет обозначено, что данная информация является конфиденциальной, то DLP-система должна отреагировать соответственно.

6) **Тезаурусы категорий**. Для каждой из категорий информации до начала работы системы должен быть создан специальный словарь терминов (групп терминов) – так называемый тезаурус категории, который должен определить возможность отнесения текста к данной категории. Соответственно, если по тезаурусу определяется, что текст не относится к какой-либо категории, дальше в алгоритме категория не рассматривается. Тезаурусы также используются при создании онтологии анализируемого текста.

Одна из основных особенностей предлагаемого метода – совместное использование

онтологий и тезаурусов. Это необходимо для достижения основной цели метода – снижение при анализе количества ошибок и трудоемкости.

Для уменьшения вероятности появления ошибок второго рода предлагается применить тезаурусы категорий, определяющие набор категорий, к которым может быть отнесен текст. Затем, для каждой из категорий, определенных выше, проводится сравнение онтологий текста и предметной области. Эта процедура предназначена для устранения ошибок первого рода. Последовательность применения онтологических методов связана с двумя факторами:

1) Быстродействие обработки информации с использованием словарей существенно выше, чем при сравнении онтологий;

2) Использование одних только тезаурусов приводит к большому количеству ошибок первого рода, т.е. к ложным срабатываниям.

Интеграция работы DLP-системы в единый алгоритм с использованием онтологий и тезаурусов позволяет существенно увеличить скорость работы системы и снизить вероятность появления ошибок.

Рассмотрим более подробно алгоритм работы анализатора:

1) Осуществляется поиск терминов тезауруса каждой категории в анализируемом тексте. Предположим, что общее количество категорий равно k. В процессе поиска по тезаурусам установлено, что текст может соответствовать n категориям. Соответственно будем считать, что оставшимся k-n категориям текст не соответствует. Таким образом, в случае k=n уже на данном этапе алгоритм может быть остановлен и принимается решение об отсутствии конфиденциальной информации.

2) По оставшимся n категориям проводится сравнение онтологии текста и той части онтологии, которая соответствует категории с 1 до n, т.е. процедура сравнения проводится n раз. Сравнение подразумевает запросы по каждому из правил онтологии предметной области, заданного для данной категории. При обнаружении совпадений происходит сопоставление текста соответствующей категории.

3) После завершения работы (завершения анализа по всем категориям) полученные результаты отправляются на обработку.

Предлагаемый метод позволяет существенно снизить размерность задачи (а, соответственно, и ее трудоемкость) и создавать на его основе DLP-системы, эффективно использующие семантику текста для поиска в нем защищаемой информации.

Список литературы

1. Панасенко А. Технические средства контроля утечек информации, 16.09.2008 [Электронный ресурс], URL: http://www.anti-malware.ru/node/569 (дата обращения: 10.04.2014).
2. Ефременко Н. Онтологии в DLP-системах третьего поколения // Журнал «Information Security/ Информационная безопасность». – 2009. – № 4. – С.32-33.
3. Добров Б.В., Иванов В.В., Лукашевич Н.В., Соловьев В.Д. Онтологии и тезаурусы: модели, инструменты, приложения. – М.: Интернет-Университет Информационных Технологий; БИНОМ. Лаборатория знаний, 2009. – 173 с.
4. Лапшин В.А. Онтологии в компьютерных системах. – М.: Научный мир, 2010. – 224 с.
5. Черняк Л. Семантический анализ на службе // Журнал «Открытые системы». – 2010. – № 10.

Рецензенты:

Смагин А.А., д.т.н., профессор, зав. кафедрой ТТС Ульяновского госуниверситета, Ульяновский государственный университет, г. Ульяновск.

Васильев К.К., д.т.н., профессор, зав. кафедрой «Телекоммуникации» Ульяновского государственного технического университета, Ульяновский государственный технический университет, г. Ульяновск.