Санкт-Петербургский национальный исследовательский университет
информационных технологий, механики и оптики

На правах рукописи

Лапшин Сергей Владимирович

МЕТОДЫ ПОВЫШЕНИЯ ПОКАЗАТЕЛЕЙ КАЧЕСТВА ФИЛЬТРАЦИИ DLP-СИСТЕМ НА ОСНОВЕ ПРЕДМЕТНО-ОРИЕНТИРОВАННОЙ МОРФОЛОГИЧЕСКОЙ МОДЕЛИ ЕСТЕСТВЕННОГО ЯЗЫКА

Специальность 05.13.19 – Методы и системы защиты информации,
информационная безопасность

Диссертация на соискание ученой степени
кандидата технических наук

Научный руководитель
д.т.н. Лебедев И.С.

Санкт-Петербург
2014

Оглавление

Оглавление 2

Список использованных сокращений 3

Введение 4

1. Защита информационных систем от утечек информации 10

1.1. Основные модели обработки естественно-языковой информации в DLP-системах 10

1.2. Основные методы борьбы с намеренными утечками информации 37

1.3. Постановка проблемы исследования 39

1.4. Выводы 40

2. Методы обнаружения угроз ИБ на основе морфологической модели естественного языка 42

2.1 Модель угрозы утечки конфиденциальной информации, обрабатываемой в современных информационных системах организаций 42

2.2 Постановка задачи 57

2.3 Метод снижения числа ошибок первого и второго рода в морфологических анализаторах DLP-систем 63

2.4 Метод предотвращения передачи конфиденциальных ЕЯ собщений, содержащих отсутствующие в словаре парадигмы слов 68

2.5 Метод идентификации защищаемых данных в передаваемых сообщениях на основе анализа связей в объектной модели естественного языка 72

2.6 Выводы 82

3. Сравнительный анализ 84

3.1 Оценка показателей качества предложенных решений 84

3.2 Оценка применимости предложенных решений 94

3.3 Выводы 100

Заключение 104

Литература 108

Список использованных сокращений

БД – база данных

БПФ – быстрое преобразование Фурье

ВКФ – взаимокорреляционная функция

ЕЯ – естественный язык

ГЗ – грамматики зависимостей

ГОС – грамматика обобщённых составляющих

ГП – грамматические переменные

ГФС – грамматика с фазовой структурой

ИБ – информационная безопасность

ИС – информационная система

ИТ – информационные технологии

КСГ – контекстно-свободные грамматики

ЛФГ – лексико-функциональные грамматики

НСГ – грамматики непосредственно составляющих

ПС – поисковые системы

РСПГ – грамматика расширенных сетей переходов

СЗИ – средства защиты информации

УГ – унификационные грамматики

DLP – data leak prevention

BYOD – bring your own device

Введение

Актуальность темы исследований

Количество зарегистрированных утечек конфиденциальной информации увеличивается с каждым годом. Это связано как с развитием и повсеместным распространением информационных систем, применяемых для обработки данных, так и с увеличением ценности самих информационных активов компаний. На рисунке 1 показан постоянный рост числа зарегистрированных утечек конфиденциальной информации с 2006 по 2013 годы [74].

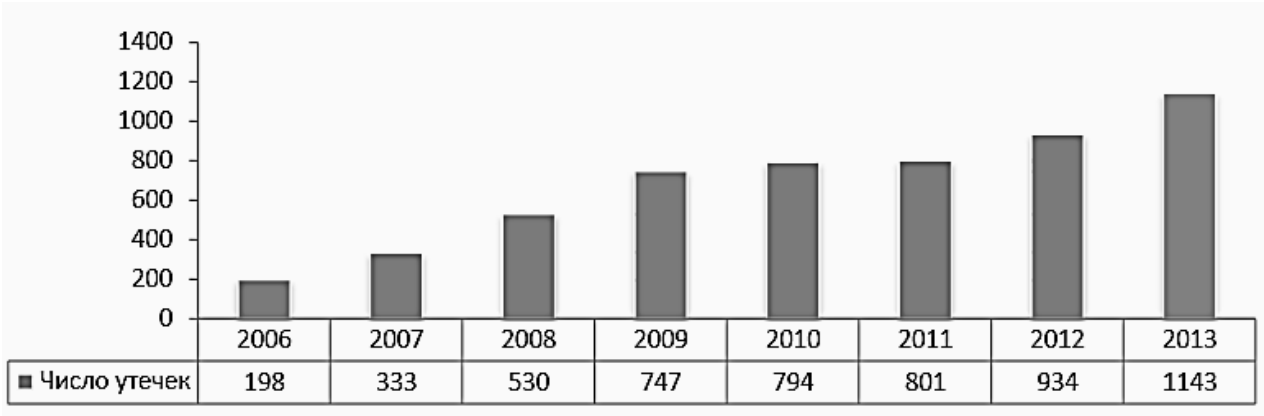

Рисунок 1. Число зарегистрированных утечек информации с 2006 по 2013 годы.

Самым действенным элементом обеспечения безопасности данных в корпоративных информационных системах остается применение технических средств защиты информации – прежде всего средств предотвращения утечек данных (Data Leak Prevention, DLP) [1].

В настоящий момент можно выделить несколько ключевых направлений развития информационных систем (ИС), которые неизбежно повлияют на идеологию DLP-систем. Во-первых, это принципиальное отличие типов информации и требований к ее защите в зависимости от специфики организаций. Даже в компаниях, относящихся к одному и тому же сегменту (банки, госорганизации, телекоммуникации), структура информационных активов неодинакова.

Во-вторых, наблюдается переход на комуникацию через разновидности «социальных сетей» с помощью мобильных устройств. [2] Это накладывает

определенный отпечаток на передаваемые сообщения: по сравнению, к примеру, с классической перепиской по e-mail, они короче, их стиль ближе к разговорной речи, а также существенно чаще встречаются специфические выражения и аббревиатуры. Анализ таких сообщений с помощью статистических методов, которые хорошо зарекомендовавали себя в поисковых задачах, затруднителен в силу специфики, которая приведена выше.

Естественно-языковые сообщения, обрабатываемые в корпоративных ИС, могут содержать защищаемую информацию как в исходном виде (так, как она хранится в виде документов и прочих носителей защищаемой информации), так и в измененном – преобразованном в другую формулировку, содержащему сокращения, специфические для отрасли компании термины и сленговые выражения и т.д.

Для решения задачи выявления DLP-системой угрозы утечки конфиденциальной информации необходимо использование лингвистических технологий, позволяющих выявить попытку передачи защищаемых данных как в исходном, так и в измененном виде.

Таким образом, повышение характеристик устойчивости обработки, полноты, точности, адекватности идентифицируемых конструкций естетсвенного языка (ЕЯ) позволяет увеличить показатели качества обнаружения угроз хищения и модификации документов, повысить показатели защищенности информации в процессе хранения и обработки и уменьшить вероятностные показатели преодоления системы защиты.

Сложность практической реализации методов автоматической обработки естественно языковых текстов и идентификации, содержащихся в них данных, на уровне семантики, существенно затрудняет достижения показателей полноты, точности вычисления текстовой информации для методов и средств пассивного и активного противодействия угрозам информационной безопасности.

Возникает противоречие между возможностями, которые предоставляют современные информационные технологии, и существующим научно-методическим и математическим обеспечением DLP-систем, реализующих

алгоритмы автоматизированной обработки текстов ЕЯ с целью выявления угроз информационной безопасности.

Следствием неразрешенности этого противоречия является необходимость разработки методов повышения показателей качества анализа естественно-языковых сообщений в DLP-системах.

Таким образом, обоснование и разработка методов повышения показателей качества обнаружения угроз утечки конфиденциальной информации за счет повышения показателей качества анализа естественно-языковых сообщений является актуальной научной задачей.

Объектом исследования являются системы предотвращения утечек информации (DLP-системы).

Предметом исследования являются методы обнаружения угрозы утечки конфиденциальной информации на основе анализа текстов ЕЯ.

Целью диссертационной работы является разработка методов повышения показателей качества выявления угрозы утечки информации DLP-системами. Для достижения указанной цели в диссертации решаются следующие **научные и технические задачи**:

1. Анализ тенденций развития корпоративных ИС с целью построения системы защиты от утечек конфиденциальных данных.

2. Анализ эффективности существующих методов анализа ЕЯ-сообщений, их применимость к современным и создаваемым СЗИ ИС.

3. Исследования проблем разработки и применения методов и средств защиты информации (DLP-систем) в процессе сбора, хранения, обработки, передачи и распространения от угрозы хищения (утечки) конфиденциальной информации.

4. Разработка методов повышения показателей качества защиты DLP-систем в выбранных направлениях.

5. Исследование характеристик СЗИ, основанных на предлагаемых методах повышения показателей качества защиты.

Научная новизна. В работе предложены методы повышения показателей качества обнаружения угрозы утечки информации за счет улучшения полноты и точности анализа текстов ЕЯ:

1. Предложен метод снижения числа ошибок первого и второго рода в морфологических анализаторах DLP-систем, отличающийся от известных использованием автоматически сформированного множества корректных полных шаблонов предложений для каждого анализируемого предложения.

2. Предложен метод предотвращения передачи конфиденциальных ЕЯ собщений, содержащих отсутствующие в словаре парадигмы слов, отличающийся от известных применением классификации по флексии основной словоформы при пополнении словаря.

3. Предложен метод идентификации защищаемых данных в передаваемых сообщениях на основе анализа связей в объектной модели естественного языка, отличающийся от известных методов использованием функции корреляции ряда связей семантических объектов.

Практическая значимость. Предложенные в работе методы позволяют повысить показатели качества обнаружения угрозы утечки информации за счет повышения точности и уменьшения вычислительной сложности анализа текстов ЕЯ, характерных для соврменных ИС. Основное внимание при этом уделено тому, что защищаемые данные могут содержаться в передаваемых сообщениях в измененной различными способами формулировке.

Метод снижения числа ошибок первого и второго рода в морфологических анализаторах DLP-систем позволяет сузить множество гипотез о морфологических характеристиках слов в передаваемом сообщении, тем самым увеличивая вероятность корректного распознавания естественноязыковых конструкций морфологическим анализатором DLP-системы.

Метод предотвращения передачи конфиденциальных ЕЯ собщений, содержащих отсутствующие в словаре парадигмы слов, позволяет автоматически получить морфологическое описание несловарного термина в анализируемом сообщении и пополнить морфологический словарь всеми его словоформими.

Благодаря этому DLP-система может более корректно анализировать характерные для современных ИС ЕЯ-сообщения. Также появляется возможность уйти от последовательного внесения в морфологический словарь всех возможных словоформ с их морфологическими характеристиками, что является необходимой, но нетипичной задачей для служб ИБ и ИТ.

Метод идентификации защищаемых данных в передаваемых сообщениях на основе анализа связей в объектной модели естественного языка позволяет уйти от вычислительно сложной задачи сравнения семантических графа передаваемого сообщения и и графа защищаемых данных за счет оценки семантических связей, учитывающей синонимию ЕЯ, и позволяющей DLP-системе с линейной сложностью по времени определять наличие защищаемых данных в передаваемых сообщениях.

Реализация результатов.

В результате реализации метода предотвращения передачи конфиденциальных ЕЯ собщений, содержащих отсутствующие в словаре парадигмы слов, получено свидетельство о регистрации ПО.

Отставлено до ясности с ведущей организацией.

Положения, выносимые на защиту.

1. Метод снижения числа ошибок первого и второго рода в морфологических анализаторах DLP-систем.
2. Метод предотвращения передачи конфиденциальных ЕЯ собщений, содержащих отсутствующие в словаре парадигмы слов.
3. Метод идентификации защищаемых данных в передаваемых сообщениях на основе анализа связей в объектной модели естественного языка.

Аппробация работы. Основные результаты работы представлялись на следующих конференциях Актуальные проблемы и технологии защиты информации, НИУ ИТМО, 2011 г., XLI научная и учебно-методическая конференция, НИУ ИТМО, 2012 г., Актуальные проблемы и технологии защиты информации, НИУ ИТМО, 2012 г., II Всероссийский конгресс молодых ученых, НИУ ИТМО, 2013 г., Всероссийская научная конференция по проблемам

информатики СПИСОК-2013, НИУ ИТМО, 2013 г., Методы и системы защиты информации. Инфорационная безопасность, НИУ ИТМО, 2014 г.

Публикации. Основные результаты работы изложены в семи публикациях, в том числе, в четырех статьях, три из которых из которых опубликованы в ведущих рецензируемых журналах, входящих в перечень ВАК общим объемом 1,56 п.л. и авторским вкладом 1 п.л..

Структура и объем диссертации. Диссертационная работа содержит введение, 3 раздела, заключение, список литературы. Объем работы составляет ___ страниц.

1. Защита информационных систем от утечек информации

1.1. Основные модели обработки естественно-языковой информации в DLP-системах

Характеристика сообщений в современных информационых системах

В настоящий момент можно выделить несколько ключевых направлений развития информационных систем (ИС), которые неизбежно повлияют на идеологию DLP-систем. Во-первых, это принципиальное отличие типов информации и требований к ее защите в зависимости от специфики организаций. Даже в компаниях, относящихся к одному и тому же сегменту (банки, госорганизации, телекоммуникации), структура информационных активов неодинакова.

Во-вторых, наблюдается переход на комуникацию через различные разновидности «социальных сетей» с помощью мобильных устройств. Это накладывает определенный отпечаток на сами передаваемые сообщения: по сравнению, к примеру, с классической перепиской по e-mail, они короче, стиль сообщений ближе к разговорной речи, а также существенно чаще встречаются специфические выражения и аббревиатуры [4]. Анализ таких сообщений с помощью статистических методов, которые хорошо зарекомендовавали себя в поисковых задачах, затруднителен в силу специфики самих сообщений, которая приведена выше.

Естественно-языковые сообщения, обрабатываемые в корпоративных ИС, могут содержать защищаемую информацию как в исходном виде (так, как она хранится в виде документов и прочих носителей защищаемой информации), так и в измененном – преобразованном в другую формулировку, содержащему сокращения, специфические для отрасли компании термины и жаргонные выражения и т.д. Для решения задачи выявления DLP-системой угрозы утечки конфиденциальной информации в этом случае необходимо использование лингвистических технологий, позволяющих выявить попытку передачи защищаемой информации как в исходном, так и в измененном виде. Таким

образом, для выявления угрозы утечки конфиденциальной информации в современных ИС DLP-системы должны гибко настраиваться с учетом возможностей естественного языка и специфики компании, в которой происходит внедрение.

Еще одна концепция, обуславливающий неизбежное изменение подходов к защите информации – инициатива BYOD (Bring Your Own Device). Более 90% сотрудников используют для работы собственные устройства, и бизнес не может игнорировать этот тренд.

Кроме того, по некоторым прогнозам налаживание деловых контактов и достижение результатов посредством обмена информацией в онлайн-пространстве неизбежно станет доминирующим видом взаимодействия в корпоративных ИС. Облачные технологии, о которых так много говорится последние пару лет, в реальности уже обеспечивают большую часть функциональности, необходимой для организации коллективной работы. В перспективе прогнозируется повсеместная адаптация технологий социальных сетей для бизнеса [2].

При анализе такого рода сообщений целесообразно использовать аналитические модели естественного языка (ЕЯ). Поэтому повышение качества анализа текстов в рамках аналитических моделей ЕЯ является необходимым условием для повышения показателй качества защиты DLP-систем.

Модели обработки ЕЯ

Как уже отмечалось выше, для решения задачи выявления DLP-системой угрозы утечки конфиденциальной информации необходимо использование лингвистических технологий, позволяющих выявить попытку передачи защищаемой информации как в исходном, так и в измененном виде. Рассмотрим основные модели ЕЯ, на которых основаны указанные лингвистические технологии.

Существующие поисковые системы (ПС) используют различные методы обработки текстов ЕЯ. В современных технологиях текстового поиска

используется не только аппарат лингвистики для анализа текстов, но и статистические методы, математическая логика и теория вероятностей, кластерный анализ, методы искусственного интеллекта, а так же технологии управления данными. Рассмотрим два основных подхода к обработке и анализу текстов ЕЯ – статистический и лингвистический (аналитический) (рис.1.1.1).

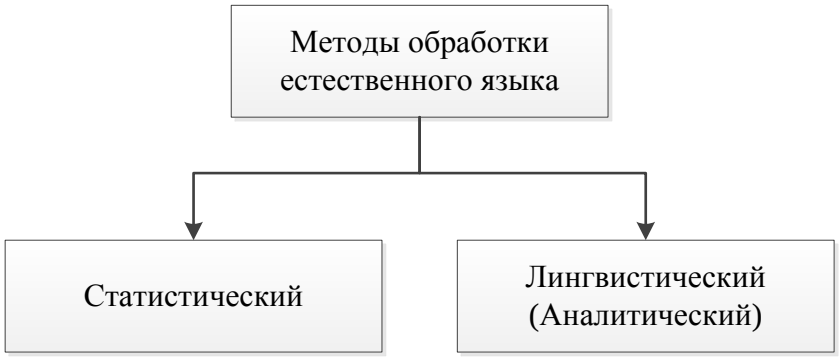

Рис. 1.1.1. Основные подходы к обработке и анализу текстов ЕЯ

В основе статистического подхода лежит предположение, что содержание текста отражается наиболее часто встречающимися словами. Суть статистического анализа заключается в подсчете количества вхождений слов в документ. Распространенным является сопоставление каждому терму t в документе некоторого неотрицательного веса. Веса термов вычисляются множеством различных способов. Самый простой из них – положить «вес» равный количеству появлений терма t в документе d, обозначается tft,d (term frequency)[43]. Этот метод взвешивания не учитывает дискриминационную силу терма. Поэтому в случае, когда доступна статистика использования термов по коллекции, лучше работает схема $tf - idf$ вычисления весов, определяемая следующим образом:

$$tf - idf\ _{i,d} = tf_{i,d} * idf_i, \text{ (1.1.1)}$$

где $idf_i = \log\frac{N}{df_i}$ – обратная документальная частота (inverse document frequency) терма t, dft - документальная частота (document frequency), определяемая как количество документов в коллекции, содержащих терм t, N –

общее количество документов в коллекции. Схема $tf - idf$ и ее модификации широко используются на практике.

Эффективным подходом, основанным на статистическом анализе, является латентно-семантическое индексирование. Латентно-семантический анализ – это теория и метод для извлечения контекстно-зависимых значений слов при помощи статистической обработки больших наборов текстовых данных [44]. Латентно-семантический анализ основывается на идее, что совокупность всех контекстов, в которых встречается и не встречается данное слово, задает множество обоюдных ограничений, которые в значительной степени позволяют определить похожесть смысловых значений слов и множеств слов между собой.

Главный недостаток статистических методов состоит в невозможности учета связности текста, а представление текста как простого множества слов недостаточно для отражения его содержания. Текст представляет набор слов, выстроенных в определенной заданной последовательности. Преодолеть этот недостаток позволяет использование лингвистических методов анализа текста.

Существуют следующие уровни лингвистического анализа: графематический, морфологический, синтаксический, семантический. Результаты работы каждого уровня используются следующим уровнем анализа в качестве входных данных (рис. 1.1.2).

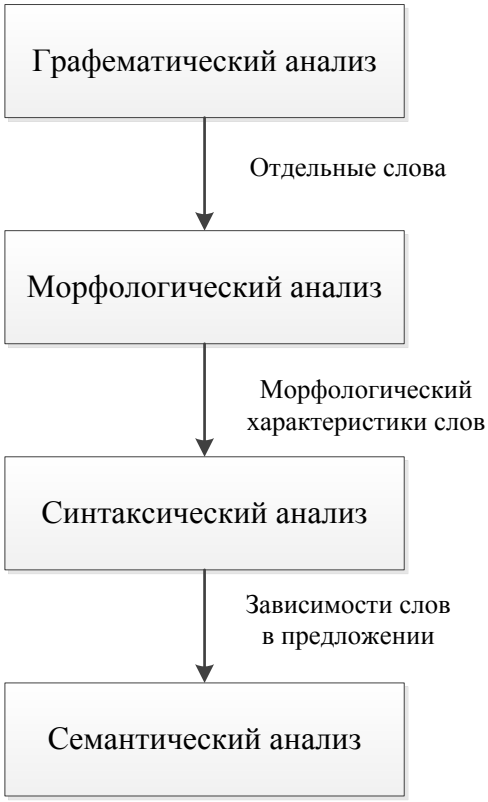

Рис. 1.1.2. Основные этапы лингвистического анализа

Целью графематического анализа является выделения элементов структуры текста: параграфов, абзацев, предложений, отдельных слов и т. д.

Целью морфологического анализа является определение морфологических характеристик слова и его основной словоформы. Особенности анализа сильно зависят от выбранного естественного языка.

Целью синтаксического анализа является определение синтаксической зависимости слов в предложении. В связи с присутствием в русском языке большого количества синтаксически омонимичных конструкций, наличием тесной связи между семантикой и синтаксисом, процедура автоматизированного синтаксического анализа текста является трудоемкой. Сложность алгоритма увеличивается экспоненциально при увеличении количества слов в предложении и числа используемых правил.

Разработки в области семантического анализа текста связаны с областью искусственного интеллекта, делающей акцент на смысловом понимании текста. В настоящее время успехи в этом направлении достаточно ограничены.

Разработанные семантические анализаторы обладают высокой вычислительной сложностью и неоднозначностью выдаваемых результатов [45].

В ходе развития информационно-поисковых систем было предложено множество моделей информационного поиска. Поскольку задачу выявления DLP-системой защищаемых данных в передаваемом сообщении можно отнести к поисковым, то далее рассмотрим основные модели информационного поиска.

Модель поиска – это сочетание следующих составляющих [46]:

1. Формат представления документов.
2. Формат представления запросов. Запрос – формализованный способ выражения информационных потребностей пользователя ПС. Для этого используется язык поисковых запросов, синтаксис которых варьируется от системы к системе.
3. Функция соответствия документа запросу. Степень соответствия запроса и найденного документа (релевантность) – субъективное понятие, поскольку результаты поиска, уместные для одного пользователя, могут быть неуместными для другого.

В различных моделях ПС вид критерия релевантности документов зависит от вида модели информационного поиска, например в моделях семантического поиска, точное вхождение слов запроса в документ не является основополагающим критерием, как, например, в теоретико-множественных моделях.

Вариации этих составляющих определяют множество реализаций систем поиска. Рассмотрим наиболее распространенные модели поиска.

Модели традиционного информационного поиска принято делить на три вида (рис.1.1.3):

1. Теоретико-множественные (булевская, нечетких множеств, расширенная булевская),
2. Алгебраические (векторная, обобщенная векторная, латентно-семантическая, нейросетевая)
3. Вероятностные

Рис. 1.1.3. Модели традиционного информационного поиска.

Булевская модель – модель поиска, опирающаяся на операции пересечения, объединения и вычитания множеств. Запросы представляются в виде булевских выражений из слов и логических операторов. Релевантными считаются документы, которые удовлетворяют булевскому выражению в запросе. Основной недостаток булевской модели заключается в непригодности для ранжирования результатов поиска.

Векторная модель – представление коллекции документов векторами из одного общего для всей коллекции векторного пространства. Документы и запросы представляются в виде векторов в N-мерном евклидовом пространстве. Вес термина в документе можно определить различными способами. Например, можно подсчитать количество употреблений терма в документе, так называемую частоту терма, – чем чаще слово встречается в документе, тем больший у него будет вес. Если терм не встречается в документе, то его вес в этом документе равен нулю.

Все термы, которые встречаются в документах обрабатываемой коллекции, можно упорядочить. В результате получится вектор, который и будет представлением данного документа в векторном пространстве.

Размерность этого вектора, как и размерность пространства, равна количеству различных термов во всей коллекции, и является одинаковой для всех документов. Релевантность в данной модели выражается через подобие векторов. Для вычисления подобия векторов используется косинусная метрика. Учитывать частотные характеристики слов предложили в 1957 году Joyce и Needham, и в 1968 году векторная модель была реализована Джерардом Солтоном (Gerard Salton (Sahlman)) в поисковой системе SMART (Salton's Magical Automatic Retriever of Text) [47]. Векторно-пространственная модель связана с расчетом массивов высокой размерности и малопригодна для обработки больших массивов данных.

В 1977 году Robertson и Sparck-Jones реализовали вероятностную модель [48]. Релевантность в этой модели рассматривается как вероятность того, что данный документ может оказаться интересным пользователю. При этом подразумевается наличие уже существующего первоначального набора релевантных документов, выбранных пользователем или полученных автоматически при каком-нибудь упрощенном предположении. Вероятность оказаться релевантным для каждого следующего документа рассчитывается на основании соотношения встречаемости терминов в релевантном наборе и в остальной, «нерелевантной» части коллекции. Вероятностная модель характеризуется низкой вычислительной масштабируемостью и необходимостью постоянного обучения системы.

Одно из перспективных направлений развития информационно-поисковых систем – построение моделей «семантического» поиска. Семантический поиск – вид автоматизированного полнотекстового информационного поиска с учетом смыслового содержания слов и словосочетаний запроса пользователя и предложений текстов проиндексированных информационных ресурсов. Семантический поиск, например, позволяет найти документы, вовсе не содержащие слов из поискового запроса, но имеющие к ней отношение. Попытки реализации семантического поиска начались в конце 20 века. В 2000 г. P. Vakkari

[49] предложил способ поиска схожих по семантике документов на основе сопоставления их лексических векторов.

Существующие системы семантического поиска

В трудах Гавриловой Т.А., Хорошевского В.Ф. [50], [51] исследуется вопрос о применении онтологического подхода для информационного поиска. Онтологии являются методами представления и обработки знаний и запросов, и предназначены для описания семантики данных для некоторой предметной области и решения проблемы несовместимости и противоречивости понятий.

Онтологии обладают собственными средствами обработки (логического вывода), соответствующими задачам семантической обработки информации. Поэтому онтологии получили широкое распространение в решении проблем представления знаний и инженерии знаний, семантической интеграции информационных ресурсов, информационного поиска и т.д.

Определение онтологии дано в работе Gruber T.R «A Translation Approach to Portable Ontology Specifications» [52]: эксплицитная, т.е. явная спецификация концептуализации, где в качестве концептуализации выступает описание множества объектов и связей между ними.

В работе Wielinga B., Schreiber A.T., Jansweijer [53], сделана попытка дать математические определения понятий «модель концептуализации предметной области», «база знаний предметной области» и «модель онтологии предметной области».

Онтология определяет общий словарь для ученых, которым нужно совместно использовать информацию в предметной области. Она включает машинно-интерпретируемые формулировки основных понятий предметной области и отношения между ними.

В России информационно-поисковая система с использованием онтологии была впервые реализована авторами Добров Б.В., Лукашевич Н.В., Сыромятников С.В., Загоруйко Н.Г. в информационно-поисковой системе УИС «РОССИЯ» (Университетская информационная система). Поступающие на вход

информационной системы потоки документов подвергаются автоматической лингвистической обработке, включающей в себя следующие этапы: морфологический анализ, терминологический анализ, рубрицирование, аннотирование [54]. Терминологический анализ реализован на основе Тезауруса по общественно-политической тематике. На базе Тезауруса осуществляется автоматическое концептуальное индексирование входящего потока текстов и производится процедура разрешения многозначных терминов.

Основная проблема при реализации применении онтологического подхода – отсутствие достаточно больших и качественных онтологий предметных областей, особенно на русском языке.

Осипов Г.С. и соавторы предложили собственную модель семантического поиска, реализовав ее в информационно-поисковой системе «Exactus», в которой объединены статистические и лингвистические методы поиска. Из статистических характеристик текста Exactus учитывает TF*IDF веса термов и значимость фрагментов текстов (на основе HTML-разметки документов). Лингвистическая составляющая - значения синтаксем (минимальных семантикосинтаксических единиц текста) и их сочетаемость в конкретном предложении [55].

В теории коммуникативной грамматики [56] русского языка опровергается традиционное противопоставление синтаксиса семантике, которое предполагает разделение знаний о законах формирования связной речи на два уровня: знания о форме (синтаксис) и знания о значении (семантика).

Основополагающая идея коммуникативной грамматики заключается в том, что синтаксис должен изучать именно осмысленную речь, а синтаксические правила должны учитывать категориальные значения слов, чтобы иметь возможность определять обобщенные значения любой синтаксической конструкции – от слова до словосочетания и простого предложения. Очевидно, что одних морфологических характеристик недостаточно, чтобы слово стало конструктивной единицей синтаксиса. Слово-лексема еще не является синтаксической единицей, слово – единица лексики, а в разных его формах могут

реализоваться или актуализироваться разные стороны его общего значения. Таким образом, решающую роль здесь играет обобщенное значение, то есть категориально-семантический класс слова. Обобщенное значение определяет синтаксические возможности слова и способы его функционирования. Формируя и изучая связную речь, синтаксис имеет дело с осмысленными единицами, несущими свой не индивидуально-лексический, а обобщенный, категориальный смысл в конструкциях разной степени сложности. Эти единицы характеризуются всегда взаимодействием морфологических, семантических и функциональных признаков. Эти единицы получили название синтаксем. Важно подчеркнуть, что семантическое значение складывается в результате соединения категориального значения и морфологической формы, реализуется в определенной синтаксической позиции. Рассмотрение слова изолированно, в отрыве от текста, не позволяет установить синтаксическое значение, а следовательно – осуществлять семантический поиск [56].

Методы семантического поиска в информационно-поисковой системе «Exactus» применяются к обработке текстов запросов пользователей и возвращаемых документов. Семантическая обработка включает в себя построение семантического поискового образа запроса, построение семантического образа документов и сравнение получившихся образов. В результате вычисляются дополнительные виды релевантности, позволяющие фильтровать документы, не соответствующие поисковому запросу в указанном понимании, т.е. отбирать только те тексты, в которых семантическое значение синтаксемы совпадает с ее семантическим значением в запросе (что невозможно в обычных статистических методах).

Итак, приведенные традиционные модели поисковых систем изначально предполагали рассмотрение документов как множества отдельных слов, не зависящих друг от друга. Вероятностная модель характеризуется низкой вычислительной масштабируемостью, необходимостью постоянного обучения системы. Наиболее распространенными являются алгебраические теоретико-множественные модели, т.к. их практическая эффективность обычно выше.

Следует отметить, что предлагаемые в последнее время новые реализации проектов информационного поиска зачастую являются гибридными моделями и обладают свойствами моделей разных классов. Одно из перспективных направлений развития информационнопоисковых систем – построение моделей семантического поиска, основная задача которых заключается в анализе текста, т.е. извлечение смысла из текста и отображение его в формальную модель, которая позволяет находить смысловую близость двух текстов. Стоит признать, что потенциал у таких систем действительно большой, однако в настоящее время реализованы далеко не все возможные семантические технологии. По сути, сейчас они только помогают выделить ключевые слова из фраз, построенных на естественном языке и подобрать дополнительные словоформы для составления корректного поискового запроса. Данное направление методов поиска требует развития [42].

Модели представления ЕЯ

Существует два основных типа моделей ЕЯ: порождающие и аналитические. Для порождающей модели исходным пунктом является некоторая грамматика, а объектом исследования – язык, порождаемый этой грамматикой. В аналитической модели – наоборот, исходным материалом служит язык, т.е. некоторая совокупность предложений, а целью исследования является определение структуры этих предложений, их составных элементов, а также определение отношений между этими элементами.

Грамматика ЕЯ как раздел науки о языке, кроме морфологии, изучает синтаксис. Синтаксис рассматривает правила построения и отдельные разновидности словосочетаний и предложений. В синтаксисе выделяют две части:

1. Учение о словосочетании – выявляет типы синтаксических отношений между словами и разновидности подчинительных словосочетаний;
2. Учение о предложении – излагает законы построения простых и сложных предложений.

К настоящему времени наиболее разработаны синтаксические грамматики, являющиеся контекстно-свободными грамматиками (КСГ). Ниже представлены семейства синтаксических грамматик.

Цепочечные грамматики фиксируют порядок следования элементов, т.е. линейные структуры предложения, задавая их в терминах грамматических классов (артикль + существительное + предлог) или в терминах функциональных элементов (подлежащее + сказуемое). Эти грамматики реализуются на ЭВМ в виде грамматик с конечным числом состояний.

Грамматики составляющих (или грамматики непосредственно составляющих – НСГ) фиксируют лингвистическую информацию о группировке грамматических элементов, например: именная группа – состоит из существительного, артикля, прилагательного и др. модификаторов; предложная группа - состоит из предлога и именной группы и т.д. до уровня предложения. Грамматики строятся как набор правил подстановки (продукций).

Грамматики зависимостей (ГЗ) задают иерархию отношений элементов предложения (главное слово определяет форму зависимых). ГЗ основаны на разделении слов в предложении с введением иерархии. Главным в предложении является глагол в личной форме, т.к. он определяет число и характер зависимых существительных.

Несмотря на разнообразие КС-грамматик для описания языка, они не позволяют в полной мере описывать ЕЯ, т.к. не учитывают контекстные условия, сопряжённые с пониманием фраз ЕЯ. Но строить контекстно-зависимые грамматики нецелесообразно, т.к. во-первых, для них не всегда возможно построить алгоритм, их обрабатывающий, а во-вторых, контекст не всегда имеет своё лингвистическое выражение. В связи с этим были предприняты попытки модифицировать существующие КСГ. Так, Хомский предложил дополнить КС-грамматики системой трансформационных правил, работающих с деревьями составляющих. Однако эта идея не получила распространение. Другое направление заключается в том, чтобы использовать контекстно-зависимые

правила. Идея этого метода состоит в том, что правила продукций переписываются так:

A[a] –> B[b] ... C[c], где малыми буквами обозначены условия, тесты, инструкции и т.д.

В грамматике обобщённых составляющих (ГОС) введены метаправила, являющиеся обобщением закономерностей правил КСГ.

В грамматиках расширенных сетей переходов (РСП) предусмотрены тесты и условия перехода к дугам, а также инструкции, которые следует выполнить при проходе по данной дуге. В разных модификациях РСП дугам приписываются веса, тогда задача состоит в поиске минимального по стоимости пути.

Разновидностью РСПГ являются каскадные РСПГ. Каскад – это РСП, снабжённая действием «transmit». Это действие вызывает остановку процесса в данном каскаде, запоминание информации о текущем состоянии системы в стеке и переход к более глубокому каскаду с последующим возвратом в исходное состояние.

Другим видом расширенных КСГ являются лексико-функциональные грамматики (ЛФГ). В них трансформационные правила отделены от самих правил подстановки и решаются как автономные уравнения.

В процессе работы с грамматиками удобно пользоваться граф-схемами, которые позволяют обобщить и представить синтаксические структу ры в удобном виде.

Унификационные грамматики (УГ) способны воплощать грамматики различных видов, в частности они позволяют выйти за рамки синтаксического анализа на семантический уровень. Данные грамматики содержат четыре компонента: пакет унификаций, интерпретатор для правил и лексических описаний, программы обработки направленных графов, анализатор с помощью граф-схемы. УГ объединяют грамматические правила со словарными описаниями, синтаксические валентности с семантическими [73].

Рассмотрим подробнее возможность применения контекстно-зависимых грамматик Хомского, теоретико-множественной модели ЕЯ и морфологических моделей ЕЯ в морфологических анализаторах DLP-систем.

Контекстно-зависимые грамматики Хомского

Основоположником применения формальных грамматик к описанию ЕЯ можно считать Ноама Хомского. Он первый в явном виде сформулировал задачу превращения лингвистики в точную науку путем построения формализуемых моделей языка, которые могут быть реализованы на компьютере, назвав это генеративной парадигмой [73]. Рассмотрим возможность применения моделей ЕЯ на основе грамматик Хомского в морфологических анализаторах DLP-систем.

Введем следующие обозначения:

A – алфавит

L – язык

G – грамматика

Совокупность строк (или предложений) называется языком. Формально, язык L над алфавитом A – это множество строк в $A^* = \bigcup_{n=0}^{\infty} A^n$, поэтому $L \in A^*$. Следовательно, операции над строками индуцируют операции на языках. Отсюда получаем L^+ и L^* следующим образом:

а) $L^0 = \{\Lambda\}$, где $\Lambda -$ пустая строка;

б) если L_i и L_j – языки, то $L_i L_j = \{xy: x \in L_i, y \in L_j\}$;

в) $L^n = L^{n-1}L, n \in N$;

г) $L^+ = \bigcup_{n \geq 1} L^n$

д) $L^* = \bigcup_{n \geq 0} L^n$

Обозначим через $L(G)$ язык, порожденный грамматикой G. Тогда алгоритм проверки вхождения $\alpha \in L(G)$ называется грамматическим разбором; он использует α и G.

Грамматикой с фазовой структурой (ГФС) G называется алгебраическая структура, содержащая из упорядоченной четверки (N, T, P, S), где:

а) N и T – непустые конечные алфавиты нетерминальных и терминальных символов соответственно, таких что $N \cap S = \emptyset$;

б) P – конечное множество продукций, $P \in V^+ \times V^*$, где $V = N \cup T$ называется словарем G;

в) $S \in N$ называется начальным символом или источником.

Предполагая, что символ $\rightarrow$ не содержится в V, соотношение $(\alpha, \beta) \in P$ обычно записывают в виде $\alpha \rightarrow \beta$.

Понятие продукции, которую также называют правилом преобразования, должно давать возможность заменять одну строку символов другой. Терминальные символы обычно расматриваются как неизменяемые символы.

Иерархия грамматик Хомского определяется следующим образом. Пусть $G = (N, T, P, S)$ является ГФС, описанной выше. Такую грамматику называют грамматикой Хомского типа 0. Если все элементы P получаются из формы $\alpha \rightarrow \beta$, где $\alpha = \gamma_1 x \gamma_2$, а $\beta = \gamma_1 \delta \gamma_2$, $\gamma_1, \gamma_2 \in V^*$, $x \in N$, $\delta \in V^+$, то говорят что G является контекстно-зависимой грамматикой, или грамматикой Хомского типа 1. Здесь строки γ_1 и γ_2 могут рассматриваться как контекст, в котором x может заменяться посредством δ.

Существует и другое ограничение для гамматики Хомского типа 1: в каждой продукции α и β должны быть такими, чтобы $1 \leq |\alpha| \leq |\beta|$. Эти определения эквивалентны.

Если подстановки могут быть выполнены без рассмотрения контекстов, тогда можно заменить «контексты» γ_1 и γ_2 пустой строкой Λ и получить более слабое ограничение: если $x \rightarrow \beta \in P$, то $x \in N$ и $\delta \in V^+$. Этому ограничению удовлетворяют грамматики Хомского типа 2. Наконец, если P состоит только из продукций вида $x \rightarrow \beta$, где $x \in N$ и $\delta \in T \cup TN$ (так, что правая часть является или единичным терминалом, или единичным терминалом, за которым следует единичный нетерминал), то говорят что G является грамматикой Хомского типа 3.

Часто бывает полезно использовать более общие формы внутри множества продукций, хотя формально это не разрешается. Удобно включать пустую строку

Λ в качестве правой части любой продукции. Такие Λ-продукции крайне необходимы с общей точки зрения, если только $\Lambda \in L$. В этом случае мы можем добавить $S \to \Lambda$ к P при условии, что S не втречается в правой части любой продукции. Однако в некоторых случаях необходимо разрешать также и более общие Λ-продукции. Чтобы различать грамматики Хомского и те грамматики, в которых разрешаются Λ-продукции, вводятся расширеные версии грамматик Хомского типа 2 и 3 – контестно-свободные и регулярные грамматики соответственно.

Языки, порожденные каким-либо из этих типов грамматик, имеют аналогичные названия. Так, структурная грамматика порождает структурный язык, структурная грамматика Хомского типа 1 – язык Хомского типа 1, контекстно-свободная грамматика – контекстно-свободный язык, а регулярная грамматика порождает регулярный язык [5].

На практике конеткстно-свободные грамматики используются достаточно широко, однако в чистом виде непригодны для полноценного анализа, так как их мощности недостаточно для работы с синтаксическими явлениями, характеризующимися наличием нелокальных связей, как, например, разрывные составляющие и эллипсис [6].

Контекстно-зависимые грамматики, будучи несколько более мощными чем КС-грамматики, позволяют анализировать большинство типов синтаксически релевантных нелокальных связей. Наиболее распространенными формальными системами данного класса являются древоприсоединительные грамматики [7].

Контекстно-зависимые грамматики позволяют определять языковые структуры, не охваченные контекстно-независимыми грамматиками, но их практическое применение при создании анализаторов для DLP-систем сопряжено с некоторыми сложностями следующего характера.

1. При использовании контекстно-зависимых грамматик резко возрастает количество правил и нетерминальных символов.
2. В контекстно-зависимых грамматиках размывается структура фраз языка, столь ясно представимая с помощью контекстно-независимых правил.

3. При попытке описать более сложные соглашения и обеспечить семантическую согласованность самой грамматики теряются многие преимущества разделения синтаксическою и семантического компонентов языка.
4. Контекстно-зависимые грамматики не решают проблемы построения семантического представления значения текста. Анализатор, который просто принимает или отвергает предложение, никому не нужен. Он должен возвращать эффективное представление семантического значения предложений [21].

Теоретико-множественная модель ЕЯ

Наиболее развитой из формализованных систем для семантики является заимствованная из математики теоретико-множественная семантика, использующая аппарат математической логики. [22] Рассмотрим возможность применения этой модели в морфологических анализаторах DLP-систем.

Назовем конечное множество G словарем. Элементами G будут слова. Рассмотрим свободную полугруппу T на G, а именно, множество всех конечных последовательностей слов с определенной на нем ассоциативной и некоммутативной бинарной операцией конкатенации. Поскольку мы рассматриваем только конечные последовательности, мы не будем особо оговаривать их конечность. Последовательность слов мы будем называть также последовательностью (цепочкой) над G. Нулевая последовательность, обозначим ее θ, – это последовательность такая, что $\theta x = x\theta = x$ для каждой последовательности x. Если это специально не оговорено, θ не принадлежит G.

Подмножество $F \in T$ назовем языком над G. Полугруппа T в целом будет в этом случае полным или универсальным языком над G.

Порождающей грамматикой языка F является конечное множество правил (называемых грамматическими правилами), задающее все последовательности из F (и только эти последовательности) и сопоставляющее каждой последовательности из F описание ее структуры, которое определяет, из каких

элементов состоит последовательность (их порядок, иерархию и взаимозависимость), а также содержит другую грамматическую информацию, необходимую для того, чтобы определить, как используется и понимается данная последовательность. Следует отметить, что описание структуры возникает в результате применения грамматических правил, т.е. структура последовательности зависит не только от языка, но и от граматики, каторой он задается.

Аналитическая грамматика языка F исходит из предположения, что F уже задан, и ставит перед собой цель получить описание последовательностей, принадлежащих множеству последовательностей из F, «изнутри», т. е. как описание отношений между словами и подпоследовательностями в зависимости от их позиции в последовательности. Такая точка зрения тесно связана с традицией структурной лингвистики, в особенности с традицией так называемой дескриптивной лингвистики.

Для того, чтобы провести ясное различие между порождающей и аналитической грамматикой, рассмотрим следующий пример. Известно, что язык с конечным числом состояний может быть задан несколькими разными способами. Если используется неоднозначная грамматика, то мы можем обнаружить так называемую структурную омонимию, которая возникает в тех случаях, когда предложение имеет как бы несколько различных «конструкций». Различие грамматических структур может быть установлено с помощью грамматики с конечным числом состояний, допускающей неоднозначности, или с помощью недетерминированного конечного автомата. Такая ситуация является типичной для порождающей грамматики.

Рассмотрим теперь другую ситуацию. Назовем две последовательности x и y эквивалентными в F, если для каждой пары последовательностей u,v выполняется либо $uxv \in F,\ uyv \in F$, либо $uxv \in T - F,\ uyv \in T - F$. Основной результат, полученный Рабином и Скоттом [9] (теорема 1), а также теорема Бар-Хиллела и Шамира [8] говорят, что F является языком с конечным числом состояний тогда и только тогда, когда в нем существует конечное число классов эквивалентности.

Такое определение языков с конечным числом состояний, которое предполагает знание только их внутренней структуры, является основным принципом аналитической грамматики.

Приведенный выше пример показывает не только различие, но также и тесную связь между этими двумя типами грамматик. Каждая из них дополняет описание, задаваемое другой.

Полезность аналитического подхода к языкам может быть показана на следующем факте. В том случае, если G конечно, универсальный язык T образует перечислимое множество и, следовательно, множество всех языков над G неперечислнмо. С другой стороны, как это показано в [10], множество всех порождающих грамматик над G (точнее, множество всех грамматик непосредственных составляющих над G) перечислимо. Следовательно, существует такое неперечислимое множество L языков над G, что $L' \in L$, и для этих языков не существует порождаюшей грамматики. Для таких языков аналитическое исследование их структуры является единственно возможным методом их изучения. Аналитическое исследование может быть проведено по отношению к любому языку, если считать возможным полный просмотр любых множеств последовательностей.

Существует много проблем, относящихся к языку F, которые могут успешно изучаться без какого-либо специального уточнения структуры F, т. е. только при том предположении, что F – определенное подмножество свободной полугруппы над G такое, что для каждой последовательности над G мы можем сказать, принадлежит она F или нет. Примером одной такой проблемы может служить морфологическая омонимия. Мы можем сказать, что морфологическая омонимия слова x не больше, чем морфологическая омонимия слова y, если для каждой пары последовательностей u,v таких, что $uxv \in F$, выполняется $uyv \in F$.

Если, кроме того, обратное неверно (т. е. существуют две такие последовательности u и v такие, что $uyv \in F$, но $uxv \in T - F$), то мы можем сказать, что морфологическая омонимия x меньше, чем морфологическая омонимия y [11].

Теоретико-множественная модель является наиболее развитой из формализованных систем для семантик. Отсутствие в этой системе ориентации на конкретный тип объектов порождало надежды на то, что ее развитие даст возможность описать семантику всего языка, однако эти надежды пока не оправдались, несмотря на различные усовершенствования, такие, как модальная логика или нечеткие множества [22].

Морфологические модели естественного языка

В силу специфики анализируемых DLP-системой сообщений (небольшая длина сообщений, наличие специфических терминов, жаргонных выражений, аббревиатур и т.д.) эффективный анализ с помощью статистических методов, которые хорошо зарекомендовавали себя в поисковых задачах, затруднителен. Для решения задачи выявления DLP-системой угрозы утечки конфиденциальной информации в этом случае необходимо использование лингвистических технологий, основанных на морфологических (аналитических) моделях ЕЯ.

В теоретических работах строятся многоуровневые формальные модели морфологии, в большинстве своем, предназначенные для синтеза. Такие модели морфологического синтеза подразумевают наличие больших словарей со сложной структурой. Они описывают широкий круг морфологических явлений. Многие компоненты этих моделей избыточны для задач машинного анализа (фонетическая реализация слова, акцентная парадигма, большое число словообразовательных аффиксов) [17].

Как уже показано выше, последним этапом морфологического анализа является семантический анализ. Обшей задачей теоретической семантики считается моделирование владения языком, под которым понимается «способность говорящего по-разному выразить одну и ту же мысль и способность слушающего установить семантическое тождество внешне различных высказываний».

Этап семантического анализа недостаточно обеспечен теорией и практикой. Одной из задач семантики является снятие лексической и структурной

неоднозначности. Для этого используется аппарат селективных ограничений, который привязан к рамкам предложений, т.е. вписывается в синтаксическую модель. Альтернативные подходы развивались на ранних этапах развития ПЯ-систем. Это тезаурусиый подход (М. Мастерман) и корреляционный анализ (С. Чеккато).

Прямой переход от поверхностных синтаксических деревьев к соответствующим представлениям смысла слишком сложен вследствие большой синонимичности языка. Поэтому в последнее время в качестве некоторого переходного элемента между синтаксисом и семантикой стали использовать глубинные синтаксические структуры (ГСС). Для описания ГСС используют т.н. $\Delta-$ грамматики (Гладкий), работающие с деревьями зависимостей.

Распространённым типом реализации семантического этапа является построение надежных грамматик. В основе грамматики лежит понятие глубинного или семантического падежа. Падежная рамка глагола является расширением понятия валентность: это набор смысловых отношений, которые могут (обязательно или факультативно) сопровождать глагол и его вариации в тексте (например: агент, адресат, цель и др.). В пределах одного языка одни и тот же глубинный падеж реализуется разными поверхностными предложно-падежными формами.

Результатом этапа семантического анализа является семантическая структура, соответствующая предложению ЕЯ [73].

Существующие в настоящее время морфологические модели различаются в основном

по следующим параметрам.

Во-первых, морфологические модели отличаются по результатам работы основанных на них морфологических анализаторов. На вход морфологический анализатор получает словоформу некоторого ЕЯ, а на выходе может выдавать все значения грамматических характеристик (род, число, падеж, вид, лицо и т.п.) заданной словоформы, а может просто отвечать на вопрос, принадлежит ли

заданная словоформа некоторому ЕЯ или нет (в этом случае морфологические анализаторы называют акцепторами).

Во-вторых, морфологические модели могут ориентироваться на полное покрытие лексики (т.е. все лексемы, которые могут обрабатывать программы морфологического уровня находятся в базе данных) или частичное покрытие лексики (морфологическая модель учитывает возможность появления лексемы, не занесенной в базу данных).

В-третьих, морфологические модели различаются по способу представления и членения словоформ. Существует два основных способа представления лексем.

1) В базе данных хранятся все словоформы всех лексем (возможно, с набором их грамматических характеристик), и каким-то образом определяются словоформы, принадлежащие одной лексеме. Такой способ представления лексем удобен и эффективен для малофлективных языков, в которых различные грамматические категории реализуются, в основном, не с помощью вариации флексий, а некоторым грамматическим способом, например, с помощью предлогов. К малофлективным языкам относится, например, английский язык.

2) В базе данных хранятся основы лексем и списки флексий (возможно, с приписанными им значениями грамматических характеристик), которые присоединяются к основе для получения какой-либо словоформы. Такой способ представления лексем эффективен для флективных языков, в которых различные грамматические категории реализуются путем вариации флексий. Флективным является, например, русский язык. Модели, в которых принят данный способ представления лексем подразделяются еще на две группы: в одной учитываются чисто орфографические основы и флексии, в другой – так называемые псевдоосновы (неизменяемая начальная часть слова) и псевдофлексии (варьируемая при словоизменении конечная часть слова). Выбор того или иного варианта определения основы связан, в основном, с эффективностью реализации и назначением морфологического компонента в целом.

В любой морфологической модели, учитывающей значения грамматических характеристик лексем, с каждой лексемой связаны: синтаксический класс (часть

речи), словоизменительный (парадигматический) класс и значения грамматических категорий, или грамматических переменных (ГП), соответствующих синтаксическому классу. Различаются свободные и связанные ГП. Связанные ГП – ГП, присущие лексеме в целом (всем ее словоформам), например, одушевленность и род для существительных. Свободные ГП – совокупность ГП, по которым лексема изменяется, например, число и падеж для существительных.

Иногда в морфологических моделях выделяются синтаксические подклассы лексем, имеющие определенные морфологические и/или синтаксические особенности. Например, в русском языке в классе прилагательных можно выделить местоименные прилагательные («который»), притяжательные прилагательные («дядин»), порядковые числительные («второй») [12]. Как показывает практика, такие особенности приводят к некоторым проблемам при оценке эффективности работы морфологических анализаторов, поскольку различные морфологичекие словари содержат различные морфологические описание одних и тех-же слов [13].

В теоретической работе «Формальная модель русской морфологии» [14] дается полное описание морфологических явлений русского языка и 54 нестандартные решения для их формализации. Перечислим важные особенности данной модели:

1. Различение морфологического рода
2. Различение синтаксического рода
3. Отнесение темы глагола (‘- ов -’, ‘- у -’, ‘- а -’ и т . д .) к флексии
4. Метод описания чередований для существительных и различение для супплетивных основ
5. Выделения специальных признаков глагола, различные комбинации значени которых покрывают все возможные в русском языке способы видообразования (всего 32 комбинации);
6. Отсечение отрицания (частицы ‘ не ’) у существительных и прилагательных.

Недостатками такой модели является ее сложность:

1. Несколько уровней представления морфологической информации, специальные грамматики для перехода с одного уровень на другой
2. Избыточность грамматических признаков, часть из которых выделены в модели для описания частных случаев

Модели, которые используют словарь, способны дать более полный анализ словоформы (т.е. оперировать большим числом грамматических признаков). Степень точности такого анализа выше, по сравнению с моделями , которые не используют словаря.

На пространстве реальных текстов системы, использующие словарь, часто дают сбои. Это обусловлено тем, что не существует полных словарей. Лексика языка непрерывно пополняется – появляются новые слова. Для каждой предметной области существует своя терминология, свое подмножество лексики языка, и включить в общий словарь всю существующую терминологию невозможно. Равно как невозможно и перечислить все существующие имена и фамилии, которые имеют регулярное склонение.

Алгоритмы программ, работающих без словаря, используют вероятностно-статистические методы и лексиконы суффиксов или квази-суффиксов, основ или квази-основ, построенных эмпирически. В статье «Эмпирическое модели рование в вычислительной морфологии» [15] описана работающая модель морфологического анализа, не требующая объемных словарей основ открытых классов слов. Модель разработана в русле инженерной лингвистики. Каждой единице лексикона в данной модели приписаны все возможные грамматические характеристики словоформ, частью которой может являться данная единица. Анализ словоформы в модели построен на правилах поиска и сочетания единиц разных лексиконов, что приводит к унификации гипотез.

Такой анализ не использует возможности текстов, поступающих на вход системы. По сути, предлагаемый метод сводится к эмпирическому сжатию исходного словаря словоформ. Для этого выделяются общие цепочки букв в

множестве словоформ, и каждой цепочке букв приписываются всевозможные значения грамматических категорий этих словоформ. Эмпирическое сжатие грамматического словаря русского языка приводит к созданию большого числа разрозненных лексиконов разной структуры, каждый из которых требует отдельной процедуры считывания данных. В статье не описана технология формирования лексиконов. Данный подход к морфологическому анализ у нельзя назвать, в полной мере, бессловарным.

Похожий метод используется в работах Г.Г. Белоногова [18], где дается описание вероятностно-статистических методов для создания вспомогательных лексиконов на основе исходного корпуса текстов. Все алгоритмы такого рода имеют одни и те же недостатки :

1. Не используются точные лингвистические методы анализа
2. Большой объем лексиконов
3. Вероятностно - статистические методы плохо работают с малой выборкой.

Точность такого анализа намного ниже, чем для систем, работающих со словарем. Эти алгоритмы не позволяют выбирать уникальные грамматические характеристики , хотя в большинстве случаев позволяют постро ить общую основу или квази-основ у для множества словоформ и лемматизировать словоформу.

Наиболее свободная форма анализа была разработана в Чикагском Университете [16]. Модель позволяет путем статистической обработки большого массива текстов, анализируя частоту встречаемости последовательности символов в словоформах, выделять множество аффиксов и корневых морфем, релевантных для заданного языка. Программа работает с большинством европейских языков, включая русский. Работа проводилась в рамках научного исследовани я и не получила прикладного внедрения.

Но при этом бессловарная морфология сохраняет свою актуальность в задачах автоматического пополнения лексиконов [17].

Предметно-ориентированная морфологическая модель естественного языка

Как показано выше, аналитический (лингвистический) подход к разбору ЕЯ предполагает четыре уровня анализа текста: графематический, морфологический, синтаксический и семантический. Результатом, как правило, является представление исходного текста в виде графа, вершины которого – объекты, о которых идет речь в проанализированном тексте, и их свойства.

Формально предметно-ориентированная морфологическая модель естественного языка может быть описана следующим образом:

Пусть $O = \{ o_i \}$, $i=1,..., n$ – множество объектов, которые получены в результате семантического анализа текста ЕЯ.

Пусть $P = \{ p_j \}$, $j=1,..., m$ – множество свойств объектов O.

Пусть $L = \{ l_k \}$, $k=1,..., r$ – множество связей между объектами O и свойствами P.

Пусть $I = \{ i_j \}$, $j=1,..., n$ – множество защищаемых данных, представленных в виде фактов i_j.

Тогда результатом семантического анализа является граф G:

$$G = \{O, P, L\}. \quad (6)$$

Защищаемыми фактами i_j в таком случае являются связи объектов с их свойствами и объектов с другими объектами.

$$i_j = \{O_j, L_j, P_j + O_{Lj}\},$$

где O_j – объект, L_j – множество связей этого объекта, P_j – множество свойств этого объекта, и O_{Lj} – множество других объектов, связанных с объектом O_j.

Пример такого графа представлен на рисунке 1.1.4.

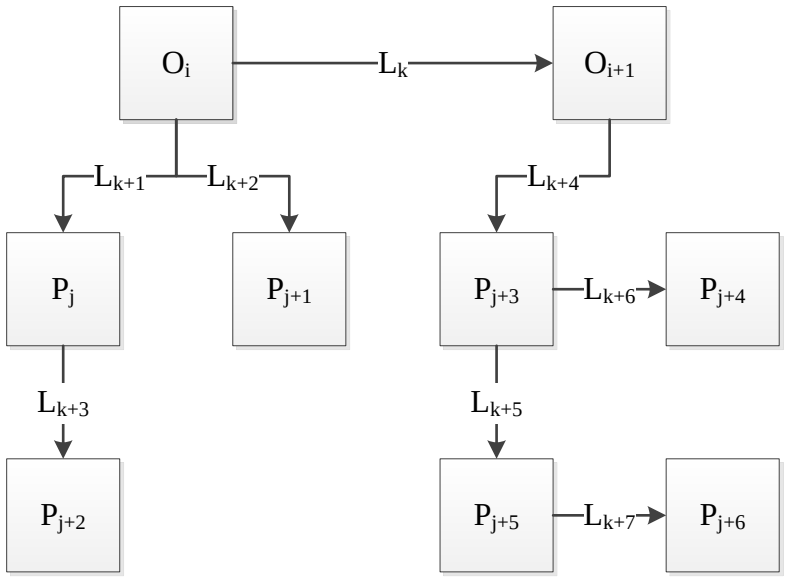

Рис 1.1.4. Результат семантического анализа текстов в рамках предметно-ориентированной морфологической модели ЕЯ

1.2.Основные методы борьбы с намеренными утечками информации

Баланс умышленных и случайных утечек в России в сравнении с общемировой картиной отличается кардинально. 77% всех российских утечек носят явно намеренный характер в то время, как общемировое распределение на протяжении десятка лет колеблется вокруг соотношения 50/50 (без учета утечек неопределенной природы).

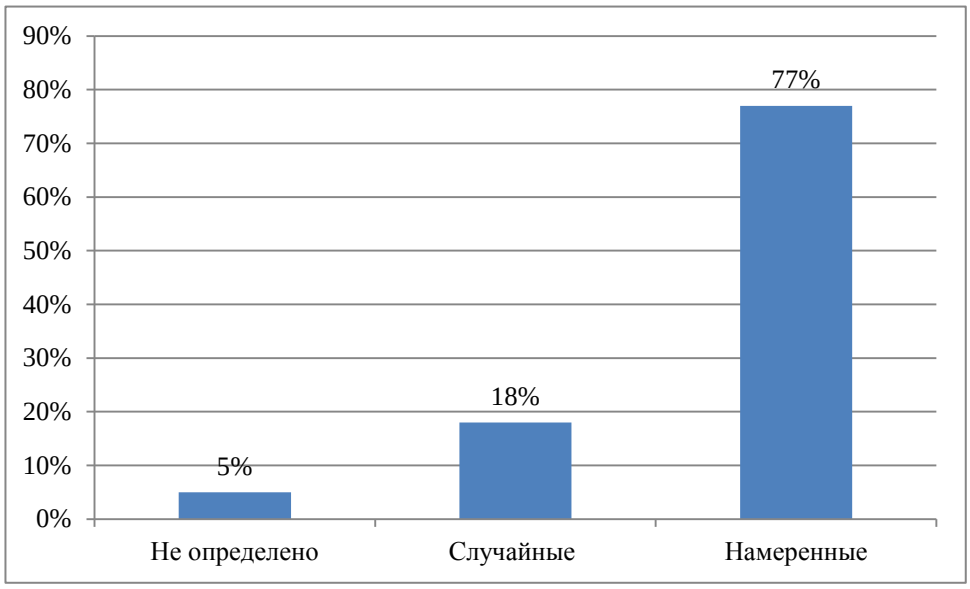

Рис. 1.2.1 Распределение утечек по причинам

Столь низкая доля случайных утечек характерна для сегментов с высоким уровнем информационной безопасности – банки, телеком-операторы. Так, в банковской сфере доля злонамеренных утечек составила 100%, т.е. утечка информации была совершена с целью наживы и перепродажи.

С точки зрения распределения по каналам утечек информации, ситуация в России не сильно отличается от общемировой. Однако есть небольшие особенности, которые следует учитывать. Во-первых, несмотря на набирающую популярность концепцию BYOD – принеси свое устройство и работай – утечек через различные мобильные устройства пока сравнительно немного –1,4% на фоне 9,6% по миру.

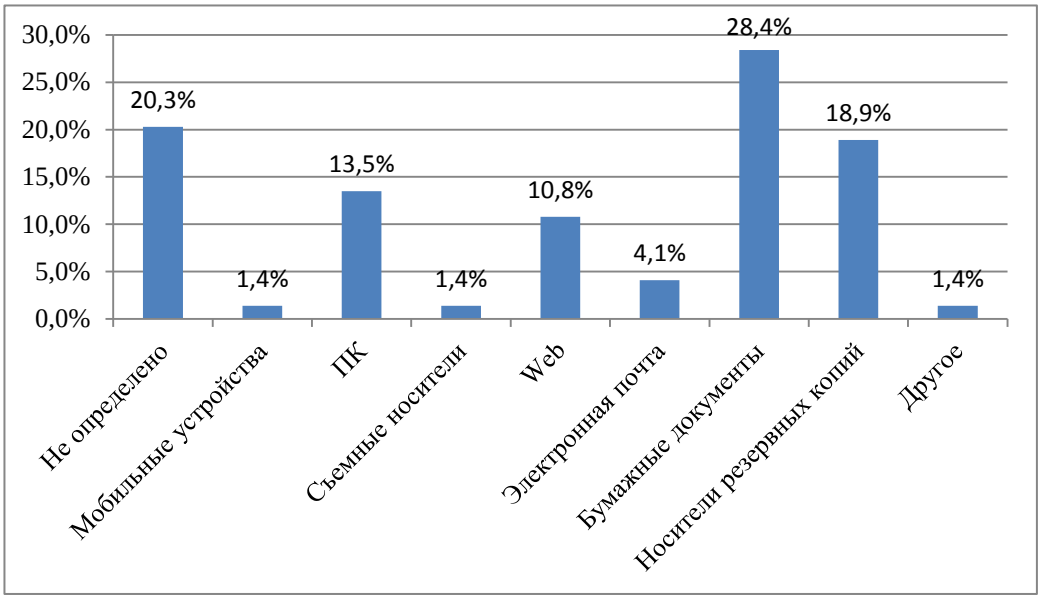

Рис. 1.2.2. Распределение утечек по каналам в России

Если принять во внимание все большую популярность средств защиты от утечек информации, в России, как и во всем мире, следует ожидать падения долей традиционных каналов утечек (где технические системы защиты наиболее эффективны). По некоторым оценкам в отношении нашей страны это справедливо в горизонте 3-5 лет [20].

Из приведенной статистики следует, что намеренные утечки информации являются серьезной проблемой. При этом в настоящий момент не существует эффективных способов противодействия утечкам такого рода. DLP-системы еще

не справляются с обнаружением этой категории утечек и предотвращают главным образом случайные, ненамеренные утечки данных [19].

Более того, гарантировать защиту информационных систем от намеренных утечек информации с помощью DLP-системы в общем случае невозможно, поскольку одним из носителей обрабатываемой в ИС информации является человеческая память, контроль которой по разным причинам невозможен.

1.3.Постановка проблемы исследования

В данной работе исследуются проблемы анализа естественно-языковых сообщений в DLP-системах с целью выявления угрозы утечки конфиденциальной информации.

Естественно-языковые сообщения, обрабатываемые в корпоративных ИС, могут содержать защищаемую информацию как в исходном виде (так, как она хранится в виде документов и прочих носителей защищаемой информации), так и в измененном – преобразованном в другую формулировку, содержащему сокращения, специфические для отрасли компании термины и жаргонные выражения и т.д.

Для решения задачи выявления DLP-системой угрозы утечки конфиденциальной информации в этом случае необходимо использование лингвистических технологий, позволяющих выявить попытку передачи защищаемой информации как в исходном, так и в измененном виде.

Таким образом, повышение характеристик устойчивости обработки, полноты, точности, адекватности идентифицируемых ЕЯ конструкций позволяет увеличить вероятность обнаружения угроз хищения и модификации документов, повысить показатели защищенности информации в процессе хранения и обработки, уменьшить вероятностные показатели преодоления системы защиты.

Сложность практической реализации методов автоматической обработки естественно языковых текстов и идентификации, содержащихся в них данных, на уровне семантики, существенно затрудняет достижения показателей полноты,

точности вычисления текстовой информации для методов и средств пассивного и активного противодействия угрозам информационной безопасности [72].

Таким образом, существует объективное противоречие между возможностями, которые предоставляют современные информационные технологии, и существующим научно-методическим и математическим обеспечением DLP-систем, реализующих алгоритмы автоматизированной обработки текста с целью выявления угроз информационной безопасности.

Следствием неразрешенности этого противоречия является объективная необходимость разработки методов повышения показателей качества анализа естественно-языковых сообщений в DLP-системах.

Таким образом, обоснование и разработка методов повышения веротности обнаружений угроз утечки конфиденциальной информации за счет повышения показателей качества анализа естественно-языковых сообщений является актуальной научной проблемой.

1.4.Выводы

1. В современных ИС для эффективной работы DLP-решений необходима доработка и разработка новых методов анализа передаваемых данных. Понятие защищенного периметра организаций, по сути, ушло в прошлое, потому системы защиты от утечек должны обеспечить безопасность данных как внутри инфраструктуры компании, так и за ее пределами. Речь идет о необходимости объединить в рамках DLP-систем технологии, позволяющие находить и контролировать принадлежащие компании данные, в том числе на просторах глобальной сети.

2. При этом, по мнению некоторых специалистов рынок DLP переживает технологический застой – в одной из лидирующих компаний-разработчиков таких систем за последние два года не зарегистрировано появления ни одной новой технологии перехвата контента [3].

3. Исследование показало, что существующие в настоящий момент модели ЕЯ нуждаются в доработке и приспособлению к нуждам лингвистических анализаторов DLP-систем.

4. В связи с этим, для повышения уровня защищенности DLP-систем необходимо разработать методы повышения качества анализа лингвистических анализаторов DLP-систем, а также выработать меры противодействия намеренным утечкам защищаемой информации.

2. Методы обнаружения угроз ИБ на основе морфологической модели естественного языка

2.1 Модель угрозы утечки конфиденциальной информации, обрабатываемой в современных информационных системах организаций

Описание защищаемой информационной системы

Современная информационная система организации состоит из множества вычислительных устройств, объединенных в сети. Общая схема такой системы представлена на рис. 2.1.1.

Рассмотрим формальную модель защищаемой информационной системы.

Пусть $I = \{ i_j \}$, $j=1,..., n$ – множество защищаемых данных, представленных в виде фактов i_j.

Пусть $D = \{ d_j \}$, $j=1,..., m$, – множество документов, содержащих, в частности, защищаемые данные: $\{ i_h, ..., i_g \} \in d_j$.

Пусть $S = \{ s_j \}$, $j=1,..., k$ – множество вычислительных устройств, обеспечивающих хранение документов D.

Пусть $T = \{ t_j \}$, $j=1,..., l$ – множество терминалов, на которых обрабатывается информация.

Пусть $U = \{ u_j \}$, $j=1,..., p$ – множество пользователей, которые обрабатывают документы D на терминалах T.

Пусть $C = \{ c_j \}$, $j=1,..., x$ – множество каналов передачи данных, по которым пользователи U могут передавать защищаемую информацию как в виде документов D, так и в виде отдельных фактов I.

Пусть $R = \{ r_j \}$, $j=1,..., y$ – множество маршрутизаторов, обеспечивающих передачу данных между устройствами хранения S, терминалами T, и внешней информационной средой по каналам C.

Пусть $A = \{ a_j \}$, $j=1,..., z$ – множество санкционированных получателей защищаемой информации. Очевидно, что в таком случае $A \cap U \neq \emptyset$, $A \cap U$ –

множество пользователей, допущенных до обработки или ознакомления с конфиденциальной информацией.

Обрабатываемая конфиденциальная информация I может постоянно храниться на множестве устройств S: на рабочих ПК пользователей, локальных серверах, а также во внешних по отношению к защищаемой организации центрах обработки данных (ЦОД) .

Обработка конфиденциальной информации I, в общем случае, также может осуществляться на любом из вычислительных устройств сети $T \cup S$. Защищаемые данные могут передаваться по каналам C как между узлами внутренней сети, так и между внутренней сетью организации и внешними ресурсами. Зачастую пользователям, которым разрешена обработка конфиденциальной информации, доступны и внешние каналы передачи информации.

Защищаемая информация может быть выведена на различные физические носители, например бумагу, оптические диски, и т.д.

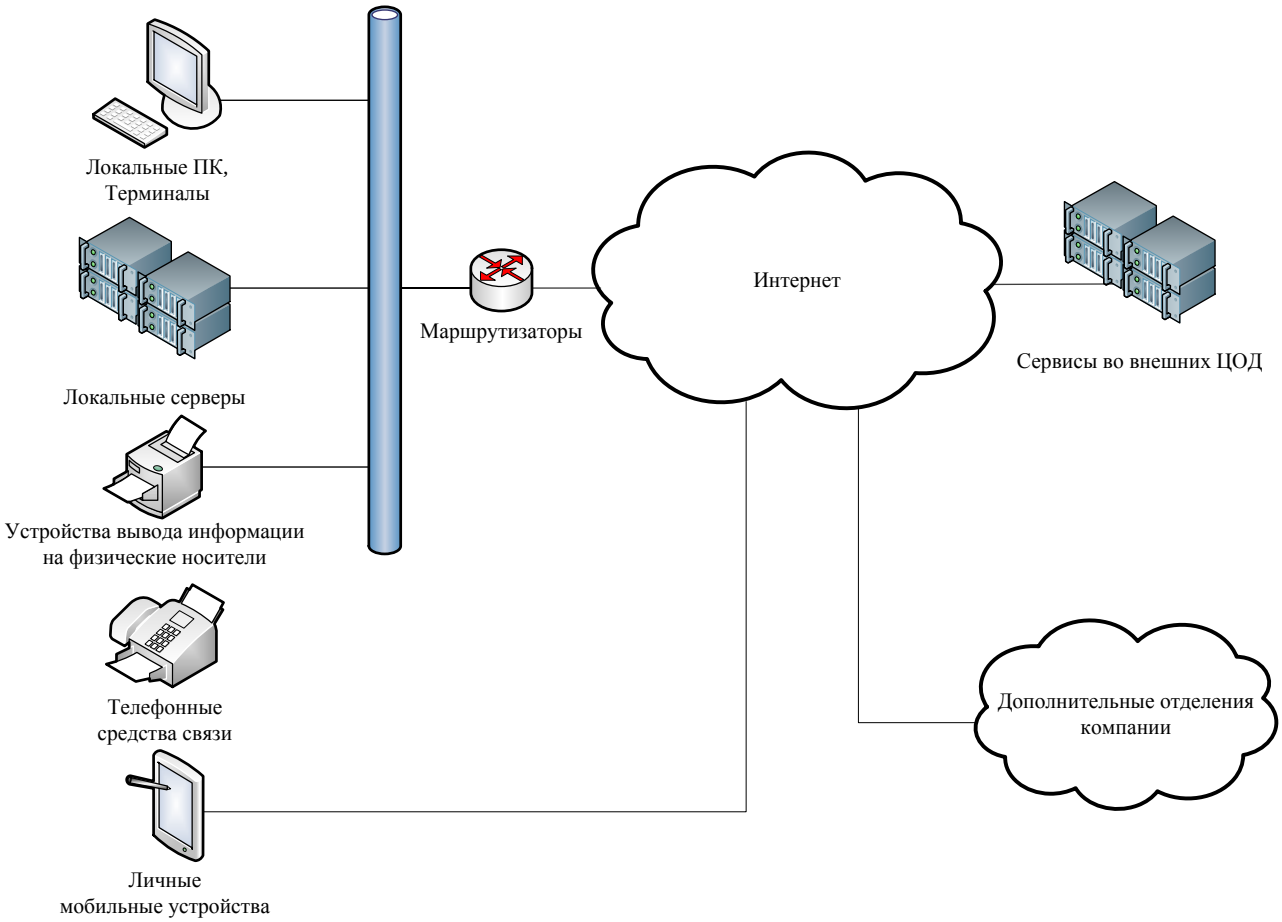

Рис. 2.1.1. Общая схема информационной системы защищаемой организации

Модель угроз конфиденциальной информации

Под угрозами конфиденциальной информации принято понимать потенциальные или реально возможные действия по отношению к информационным ресурсам, приводящие к неправомерному овладению охраняемыми сведениями.

Такими действиями являются:

- Ознакомление с конфиденциальной информацией различными путями и способами без нарушения ее целостности
- Модификация информации с частичным или значительным изменением состава и содержания сведений

- Разрушение (уничтожение) информации с целью нанесения прямого материального ущерба

В конечном итоге противоправные действия с защищаемой информацией приводят к нарушению ее конфиденциальности, целостности и доступности (Рис. 2.1.1).

Рис. 2.1.2. Угрозы информации

Отношения объекта (организация) и субъекта (конкурент, злоумышленник) в информационном процессе с противоположными интересами можно рассмотривать с позиции активности в действиях, приводящих к овладению конфиденциальными сведениями.

В общем факт получения охраняемых сведений злоумышленниками и конкурентами называют утечкой. Однако одновременно с этим в значительной части законодательных актов, законов, кодексов и официальных материалов используются и такие понятия как разлашение сведений и несанкционированный доступ к конфиденциальной информации.

Разглашение – это умышленные или неосторожные действия с конфиденциальными сведениями, приведшие к ознакомлению с ними лиц, не допущенных к ним.

Разглашение выражается в сообщении, передаче, предоставлении, пересылке, опубликовании, утере и в других формах обмена и действий с защищаемой информацией. Реализуется разглашение по формальныи и неформальным каналам

распространения информации. К формальным коммуникациям относятся деловые встречи, совещания, переговоры и тому подобные формы общения. Неформальные коммуникации включают личное общение (встречи, переписка), выставки, семинары, конференции и другие массовые мероприятия, а также средства массовой информации. Как правило, причиной разглашения конфиденциальной информации является недостаточное знание сотрудниками правил защиты конфиденциальной информации и непонимание (или недопонимание) необходимости их тщательного соблюдения.

Утечка — это бесконтрольный выход конфиденциальной информации за пределы организации или круга лиц, которым она была доверена.

Утечка информации осуществляется по различным техническим каналам. Известно, что информация вообще переносится или передается либо энергией, либо веществом. Это либо акустическая волна (звук), либо электромагнитное излучение, либо лист бумаги (написанный текст) и др. С учетом этого можно утверждать, что по физической природе возможны следующие пути переноса информации: световые лучи, звуковые волны, электромагнитные волны, материалы и вещества. Соответственно этому классифицируются и каналы утечки информации на визуально-оптические, акустические, электромагнитные и материально-вещественные. Под каналом утечки информации принято понимать физический путь от источника конфиденциальной информации к злоумышленнику, посредством которого последний может получить доступ к охраняемым сведениям. Для образования канала утечки информации необходимы определенные пространственные, энергетические и временные условия, а также наличие на стороне злоумышленника соответствующей аппаратуры приема, обработки и фиксации информации.

Несанкционированный доступ — это противоправное преднамеренное овладение конфиденциальной информацией лицом, не имеющим права доступа к охраняемым секретам.

Несанкционированный доступ к источникам конфиденциальной информации реализуется различными способами: от инициативного сотрудничества,

выражающегося в активном стремлении «продать» секреты, до использования различных средств проникновения к коммерческим секретам. Для реализации этих действий злоумышленнику приходится часто проникать на объект или создавать вблизи него специальные посты контроля и наблюдения — стационарных или в подвижном варианте, оборудованных самыми современными техническими средствами.

Если исходить из комплексного подхода к обеспечению информационной безопасности, то такое деление ориентирует на защиту информации как от разглашения, так и от утечки по техническим каналам и от несанкционированного доступа к ней со стороны конкурентов н злоумышленников.

Такой подход к классификации действий, способствующих неправомерному овладению конфиденциальной информацией, показывает многогранность угроз и многоаспектность защитных мероприятий для обеспечения информационной безопасности. [23]

Итак, утечка информации осуществляется по техническим каналам:

- визуально-оптическому
- акустическому
- электромагнитному
- материально-вещественному

Контроль технических каналов, как правило, реализуется исключительно организациоными мерами. Это связано с большой сложностью и низкой эффективностью технических средств защиты этих каналов для коммерческих организаций.

Модель угрозы утечки конфиденциальной информации

В данной работе рассматриваются утечки информации по «разрешенным» информационным каналам передачи данных *C*, которые используются для взаимодействия между узлами $T \cup S$ внутренней и внешней сети.

Под **утечкой** будем понимать передачу хотя-бы одного защищаемого факта i по каналу c так, что получатель информации $a \notin A$, т.е. получатель не входит в множество разрешенных получателей A.

Множество возможных каналов утечки информации C можно разделить на следующие группы:

- Каналы отправки почтовых сообщений
- Каналы отправки «мгновенных» сообщений
- Каналы передачи файлов
- Остальные каналы передачи данных

Такая классификация более понятна с прикладной точки зрения, однако в общем виде не несет конкретики, а при большей детализации быстро теряет актуальность.

В связи с этим, классифицировать возможные каналы утечки информации предлагается по протоколам передачи информации. Это позволяет более универсально описать большинство распространенных способов передачи информации в современных информационных системах.

Для описания возможных каналов утечек информации C предлагается использовать следующую классификацию:

- Каналы, основанные на протоколе HTTP
- Каналы, основанные на протоколе HTTPS
- Каналы, основанные на протоколе SMTP
- Каналы, основанные на протоколе FTP
- Каналы, основанные на протоколе SMB
- Каналы, основанные на протоколе ICQ
- Каналы, основанные на протоколе XMPP
- Каналы, основанные на протоколе IRC
- Каналы, основанные на протоколе Skype
- Каналы, основанные на протоколах p2p
- Остальные каналы передачи данных

Защищаемая информация может быть передана с использованием любого из вышеперечисленных протоколов. При этом и сами данные могут быть подвергнуты существенной модификации.

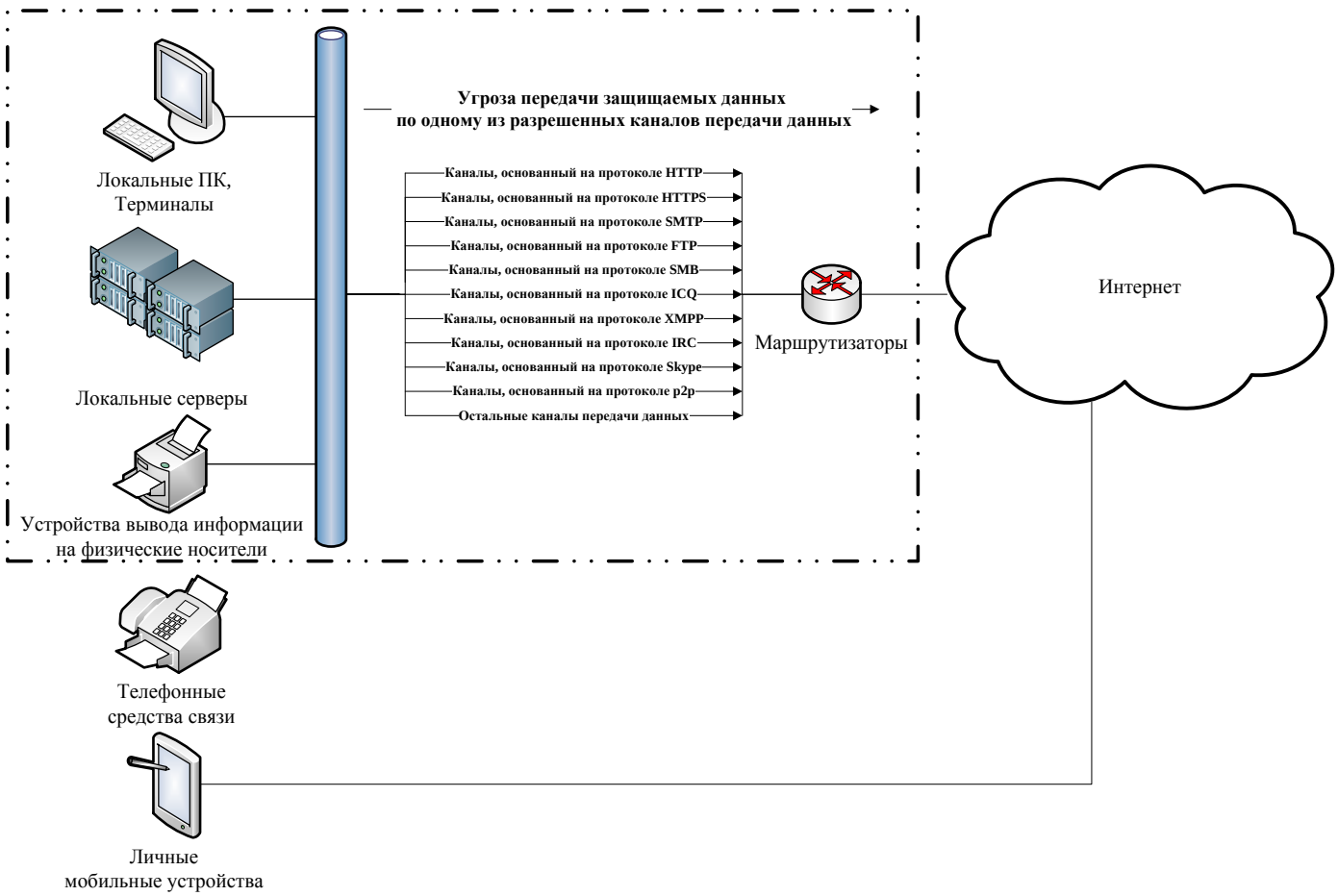

Рис. 2.1.3. Модель угрозы утечки конфиденциальной информации

Далее необходимо классифицировать сами данные, которые передаются по вышеуказанным протоколам. Физически, передаваемые данные являются сигналами (электрическими, оптическими, электромагнитными). Это самый низкий уровень описания передаваемой информации, который, как и все последующие до прикладного, неудобны для классификации в силу необходимости рассмотрения множества дополнительных структур и сущностей, которые напрямую не относятся к передаваемой информации.

В связи с этим, рассмотрим следующую классификацию передаваемых данных:

- Сообщения на естественном языке с употреблением специализированных терминов
- Бинарные объекты

Бинарные объекты, в свою очередь, могут содержать в себе сообщения на естественном языке и другие бинарные объекты. Уровень вложенности объектов при этом не ограничен.

Защищаемая информация может содержаться как в сообщениях на естественном языке, так и бинарных объектах.

Бинарные объекты могут быть модифицированы следующими способами:

- Архивирование/шифрование
- Разделение на части
- Изменение формата

Сообщения на естественном языке могут содержать в себе защищаемую информацию, которая изменена различными способами:

- Передана в другой формулировке
- Передана с грамматическими и синтаксическими ошибками
- Передана с использованием специфических терминов и оборотов
- Переведена на другой естественный язык
- Передана в другой кодировке
- Преобразована с помощью транслитерации или других способов замены символов
- Преобразована перестановкой слов
- Преобразована разделением слов на части
- Передана частями в различные моменты времени через различные каналы

Указанные способы могут сочетаться друг с другом в различных комбинациях, а также сочетаться с различными способами модификации бинарных данных.

При этом использование естественного языка приводит к тому, что при автоматическом анализе возникают дополнительные трудности. Перечислим основные из них:

- Проблема синонимии
- Проблема омонимии
- Устойчивые сочетания слов
- Морфологические вариации

Проблема синонимии. Одно понятие может быть выражено различными словами. В результате релевантные документы, в которых используются синонимы понятий, используемых в передаваемом сообщении, могут быть не обнаружены системой.

Проблема омонимии и явлений «смежных с омонимией». Грамматические омонимы – разные по значению слова, но совпадающие по написанию в отдельных грамматических формах. Это могут быть слова одной или разных частей речи. Лексические омонимы – слова одной части речи, одинаковые по звучанию и написанию, но разные по лексическому значению.

Устойчивые сочетания слов. Словосочетания могут иметь смысл отличный от смысла, который имеют слова по отдельности.

Морфологические вариации. Во многих естественных языках слова имеют несколько морфологических форм, различающихся по написанию [42].

Таким образом, утечкой является передача любого модифицированного описанными выше способами сообщения либо бинарного объкта по любому из описанных выше каналов c содержащих хотя-бы один защищаемый факт i получателю информации $a \notin A$, т.е. когда получатель не входит в множество разрешенных получателей A.

Общее описание DLP-систем

DLP-системами (от англ. Data Loss Prevention, Data Leak Prevention или Data Leakage Protection) принято называть класс решений, предназначенных для предотвращения утечек информации в информационных системах.

DLP-системы строятся на анализе потока данных, пересекающих периметр защищаемой информацонной системы. При детектировании в этом потоке конфиденциальной информации срабатывает активная компонента системы, и передача сообщения блокируется. [24]

Общая схема работы DLP-системы выглядит следующим образом:

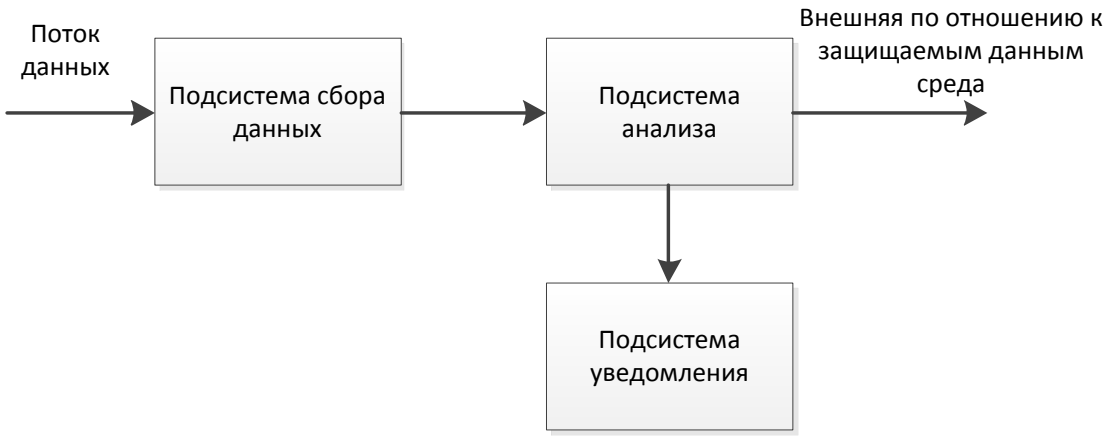

Рис. 2.1.4. Общая схема DLP-системы

Подсистема сбора данных предназначена для выделения и преобразования в единый формат анализируемой информации из всех потоков данных *C*, а также информации о поведении пользователей.

Подсистема анализа определяет возможность передачи анализируемого сообщения в среду, которая является внешней по отношению к защищаемым данным.

Подсистема уведомления предназначена для уведомления заинтересованных лиц (как минимум, офицеров безопасности) о некоторых попытках передачи защищаемой информации во внешнюю по отношению к защищаемым данным среду.

Общая схема подсистемы анализа выглядит следующим образом:

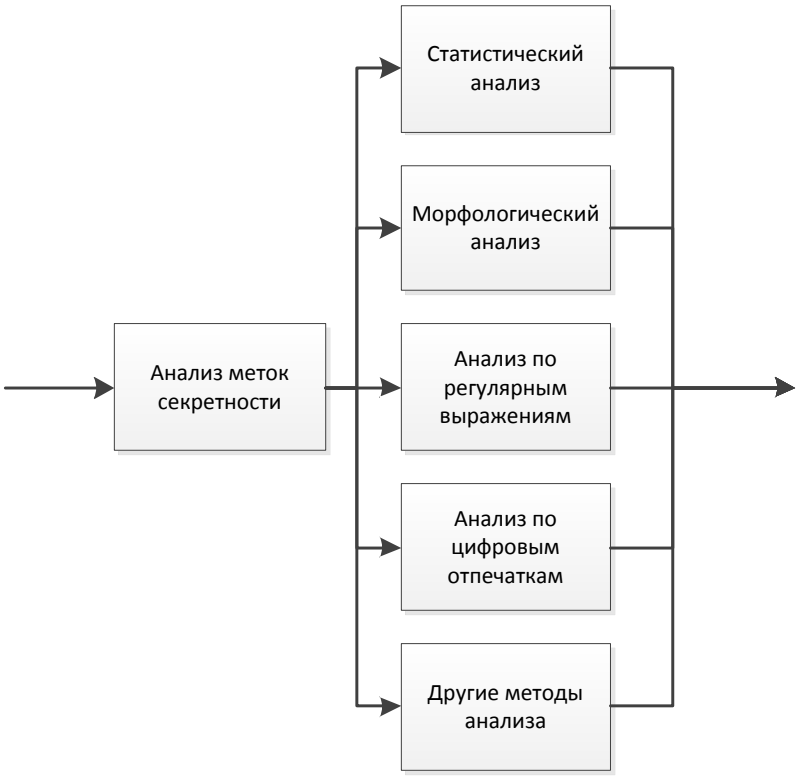

Рис. 2.1.5. Общая схема подсистемы анализа DLP-системы

Подаваемые на вход подсистеме анализа данные разделяются на сообщения, для которых метки секретности заранее определены и уже известны, и на сообщения, уровень секретности которых еще не определен.

Решение по первой группе сообщений принимается достаточно просто, на основе имеющейся метки секретности.

Для принятия решения по второй группе сообщений, для которых уровень секретности еще не определен, необходимо провести анализ. Для этого используются ряд статичтических методов анализа текстов естественного языка, методы, основанные на морфологическом анализе передаваемых сообщений, а также методы, основанные на регулярных выражениях, цифровых отпечатках и т.д.

Перечисленные методы имеют различные показатели полноты и точности поиска защищаемых данных в передаваемых сообщениях. Естественно-языковые сообщения, обрабатываемые в корпоративных ИС, могут содержать защищаемую информацию как в исходном виде (так, как она хранится в виде документов и прочих носителей защищаемой информации), так и в измененном –

преобразованном в другую формулировку, содержащему сокращения, специфические для отрасли компании термины и жаргонные выражения и т.д. Для решения задачи выявления DLP-системой угрозы утечки конфиденциальной информации в этом случае необходимо использование методов, позволяющих выявить попытку передачи защищаемой информации как в исходном, так и в измененном виде. Таким образом, для выявления угрозы утечки конфиденциальной информации в современных ИС DLP-системы наиболее целесообразно использовать аналитические методы, которые основаны на морфологическом анализе передаваемых сообщений.

Кроме того, при передаче защищаемой информации в измененной формулировке статистический метод, а также методы анализа по цифровым отпечаткам и регулярным выражениям не применимы, поскольку не учитывают синтаксические и семантические особенности ЕЯ.

В разделе 1.1 (Основные модели обработки естественно-языковой информации в DLP-системах) показано, что последний этап морфологического анализа – этап семантического анализа – недостаточно обеспечен теорией и практикой. В связи с этим для повышения показателей полноты и точности обнаружения угрозы утечки конфиденциальной информации DLP-системой необходимо развить и доработать используемые методы морфологического анализа.

Формальная модель DLP-системы

Из определения DLP-системы очевидно следует, что ее основной задачей является предотвращение утечек информации. Иными словами, DLP-система решает задачу выявления хотя бы одного из защищаемых фактов $i \in I$ в каналах передачи информации C для вынесения вердикта v о возможности дальнейшей передачи:

$$F_{DLP}(c, I) = v \quad (2.1.1)$$

Если вердикт v отрицательный, то происходит блокировка канала c, по которму совершена попытка передачи хотя-бы одного защищаемого факта i получателю информации $a \notin A$.

Подсистемы сбора данных и уведомления DLP-систем являются служебными по отношению к подсистеме анализа и не рассматриваются подробно.

Для данного исследования особый интерес представляет подсистема анализа DLP-систем. На основе вердикта, который получен от подсистемы анализа, принимается решение о возможности передачи анализируемого сообщения во внешнюю по отношению к защищаемым данным среду.

Основной задачей подсистемы анализа DLP-системы является определение содержания одного из защищаемых фактов $i \in I$ в сообщении, передаваемом по каналу c. Для этого могут использоваться уже перечисленные выше методы. Тогда функцию анализатора DLP-системы можно представить в виде объединения функций

$$F_{DLP} = F_1(c, I_{\text{норм}}) + F_2(c, I_{\text{норм}}) + \ \dots \ + F_{m-1}(c, I_{\text{норм}}) + F_m(c, I_{\text{норм}}), \quad (2.1.2)$$

где

$$I_{\text{норм}} = F_{\text{норм}}(I) \quad (2.1.3)$$

и, в зависимости от функционального наполнения DLP-системы функции F_j могут обозначать: F_1 – статистический анализ, F_2 – морфологический, F_3 – анализ по регулярным выражениям, F_4 – анализ по цифровым отпечаткам и т.д.

Функция нормализации $F_{\text{норм}}$ выполняет итеративный разбор передаваемых объектов, разделяя их на бинарные объекты и сообщения на естественном языке. В результате, полученное множество бинарных объектов отправляется на анализ в соответствии с типом каждого объекта (изображения, схемы и т.д.), а полученное множество сообщений на естественном языке передается на вход функциям F_i (2.1.2).

В предыдущем разделе уже упоминалось, что сообщения на естественном языке могут содержать в себе защищаемую информацию, которая изменена различными способами. Функция нормализации $F_{\text{норм}}$ решает проблему модификации передаваемых данных в следующих случаях:

- Передана с грамматическими и синтаксическими ошибками
- Передана в другой кодировке
- Преобразована разделением слов на части
- Преобразована с помощью транслитерации или других способов замены символов

Таким образом, последствия трех из семи способов модификации защищаемой информации могут быть определены на этапе нормализации, до начала анализа функциями F_i (2.1.2).

Важно отметить, что сам факт использования методов модификации защищаемой информации является подозрительным.

После этапа нормализации ($F_{\text{норм}}$) остается 5 способов передачи защищаемых фактов:

- Передача без изменения
- Передача в другой формулировке
- Передача с использованием специфических терминов и оборотов
- Передача на другом естественном языке
- Преобразование перестановкой слов
- Передача частями в различные моменты времени через различные каналы

Анализаторы F_i (статистический анализ, морфологический анализ, анализ по регулярным выражениям, анализ по цифровым отпечаткам и т.д.) DLP-системы предназначены для определения передачи защищаемых фактов i получателю информации $a \notin A$, т.е. когда получатель не входит в множество разрешенных получателей A. За счет этого DLP-система решает поставленную задачу предотвращения утечки конфиденциальной информации.

Перечисленные методы показывают различную эффективность при работе с разными наборами данных. Так, например, статистический анализ показывает существенно большую точность при обработке больших объемов текста на естественном языке по сравнению с обработкой коротких ЕЯ сообщений. Наиболее универсальным методом считается морфологический анализ, однако он является наиболее сложным для реализации и поддержки. Кроме того, точность современных морфологических анализаторов, в силу большой сложности задачи анализа естественного языка, не идеальна.

Таким образом, анализатор естественного языка является ключевым элементом подсистемы анализа DLP-системы. От качества его работы существенно зависит показатели качества работы всей DLP-системы, а следовательно и показатели полноты и точности обнаружения угрозы утечки конфиденциальной информации.

2.2 Постановка задачи

Целью диссертационной работы является разработка методов повышения показателей качества защиты DLP-систем. Основным компонентом DLP-системы, как уже было сказано выше, является подсистема анализа, которая формально описывается функцией анализа (фильтрации) F_{DLP} (2.1.1). Поэтому, для достижения поставленной цели необходимо максимизировать точность P_{DLP} и полноту R_{DLP} этой функции.

Применительно к функции анализа F_{DLP} термины точности P_{DLP} и полноты R_{DLP} имеют следующее значение:

$$P_{DLP} = \frac{|D_b \cap D_{leak}|}{|D_b|}, \quad (2.2.1)$$

$$R_{DLP} = \frac{|D_b \cap D_{leak}|}{|D_{leak}|}, \quad (2.2.2)$$

где D_b – множество обнаруженных угроз утечки информации, D_{leak} – множество всех утечек информции.

Иными словами, точность предотвращения утечек информации DLP-системой определяет отношение числа верно определенных утечек к числу всех

обнаруженных угроз, куда входят и ложные срабатывания системы защиты. Полнота определяет число верно определенных утечек к числу всех произошедших утечек информации.

Используя формальную модель DLP-системы, на основе формулы 2.1.1 параметры точности и полноты можно уточнить следующим образом:

$$P_{DLP} = \frac{|C_b \cap C_{leak}|}{|C_b|}, \quad (2.2.3)$$

$$R_{DLP} = \frac{|C_b \cap C_{leak}|}{|C_{leak}|}, \quad (2.2.4)$$

где C_b – множество заблокированных каналов передачи данных, C_{leak} – множество каналов, по которым совершалась попытка передачи защищаемых фактов I получателю информации a, который не входит в множество санкционированных получателей защищаемой информации, $a \notin A$.

Как уже упоминалось, существуют следующие способы модификации защищаемой информации *I*:

1. Передача с грамматическими и синтаксическими ошибками
2. Передача в другой кодировке
3. Преобразование с помощью транслитерации или других способов замены символов
4. Передача с использованием специфических терминов и оборотов
5. Передача в другой формулировке
6. Передача на другом естественный язык
7. Преобразование перестановкой слов
8. Преобразована разделением слов на части
9. Передача частями в различные моменты времени через различные каналы

Задачи определения и разрешения модификации ЕЯ сообщений первыми тремя способами являются решенными техническими задачами. Результат модификации в этом случае определяется и преобразовывается на этапе нормализации $F_{\text{норм}}$ (2.1.3).

В работе представлены методы повышения качества анализа естественноязыковых сообщений в DLP-системах при передаче защищаемых фактов без изменений, а также при использовании методов модификации 4 и 5.

Необходимость повышения точности определения морфологических характеристик

Итак, для достижения поставленной в работе цели – повышения показателей полноты и точности обнаружения DLP-системой угроз информационной безопасности – необходимо максимизировать точность P_{DLP} и полноту R_{DLP} функции F_{DLP}. Из (2.1.2) с учетом эффективности перечисленных методов следует, что для этого необходимо повысить точность P_{DLP} и полноту R_{DLP} функции морфологического анализа.

Рассмотрим функцию морфологического анализа (F_2), которая, в свою очередь, разделяется на подэтапы:

$$F_2 = F_{2k} \circ F_{2k-1} \circ \dots \circ F_{22} \circ F_{21}, \quad (2.2.5)$$

Где F_{21} – графематический анализ, F_{22} – морфологический, F_{23} – синтаксический, F_{24} – семантический и т.д.

Для повышения показателей полноты и точности обнаружения DLP-системой угроз информационной безопасности необходимо максимизировать точность P_{DLP} и полноту R_{DLP} функции F_{DLP}. Из (2.2.8) следует, что для этого, в первую очередь, необходимо повысить качество анализа на первых (нижних) уровнях – F_{21}, затем F_{22}, затем F_{23} и т.д. Для функции F_{21} (графематический анализ) уже разработаны достаточно эффективные методы, позволяющие достигнуть точности не менее 99% в определенных условиях. [25] Кроме того, при анализе коротких сообщений эта задача, в некоторых случаях, не является актуальной.

Следовательно, для повышения показателей полноты и точности обнаружения DLP-системой угроз информационной безопасности необходимо повысить эффективность работы функции F_{22}, т.е. повысить качество

определения морфологических характеристик слов в предложении. Это является первой задачей, которую необходимо решить для достижения поставленной цели работы.

Необходимость постоянной автоматической актуализации словаря морфологических описаний слов

Для достижения поставленной в работе цели – повышения показателей полноты и точности обнаружения DLP-системой угроз информационной безопасности – необходимо максимизировать точность P_{DLP} и полноту R_{DLP} функции F_{DLP}. Из (2.1.2) и (2.2.5) видно, что показатели качества P_{DLP} и полноту R_{DLP} существенно зависят от показателей качества на этапе морфологического анализа.

Основной задачей на этапе морфологического анализа (F_{22} в (2.2.5)) является определение морфологических характеристик каждого слова в передаваемом сообщении. Для этого используется морфологический словарь (иначе, словарь морфологических описаний слов).

Сообщения, циркулирующие в вычислительных сетях, обрабатываемые с целью мониторинга состояния информационной безопасности, имеют ряд особенностей. Среди них необходимо отметить небольшую длину и использование специфических выражений и аббревиатур [4]. Примером могут являться сообщения в интернет-мессенджерах или социальных сетях.

Из этого следует, что во-первых, морфологические словари DLP-системы должны помимо «общеизвестных» слов содержать специфичные для компании, где развертывается DLP-система, термины и сокращения. Во-вторых, естественный язык, особенно устная речь, содежит неологизмы. Поэтому словарь морфологических описаний слов DLP-системы также должен постоянно пополняться неологизмами. В третьих, в связи со спецификой анализируемых текстов, в таком словаре должны быть не только корректные словоформы, но и словоформы с типичными ошибками, которые допускаются людьми при написании текстов

Задача пополнения словарей морфологических описаний слов является довольно трудоемкой. Тем более, эта задача является нетипичной для служб ИБ и ИТ, которые, как правило, занимаются обслуживанием DLP-систем. Следовательно, необходимо получить способ простого, по возможности автоматического пополнения словарей морфологических описаний слов DLP-системы.

Работы в этом направлении ведутся уже достаточно давно. Большой вклад в рассматриваемом вопросе внес коллектив ЭМИ РАН [26], а также компании АОТ, Noolab, RCO и др. [27].

Морфологический словарь может быть формально описан следующим образом.

Пусть $W = \{ w_i \}$, $i=1,..., n$ – множество исходных форм слов БД СЗИ (база данных средства защиты информации).

Пусть $P = \{ p_j \}$, $j=1,..., k$ – множество парадигм исходных форм слов.

Каждому элементу множеств P и W соотвествует морфологическое описание D_j:

$$p_j \to D_j, w_j \to D_j. \quad (2.2.9)$$

Морфологическим словарем (или словарем морфологических описаний слов) будем называть совокупность множеств W, P и D с указанными между ними соответствиями (2.2.9).

Для обнаружения DLP-системой угроз информационной безопасности необходимо, чтобы морфологический словарь DLP-системы содержал все употребляемые на момент передачи сообщения слова и их словоформы: $W + P$. Кроме того, для каждого слова и словоформы должны быть указаны соответствующие им морфологические признаки D.

Как уже упоминалось, «классических» словарей для этого недостаточно, необходимое постоянное пополнение новыми словами и словоформами.

Тогда необходимо найти такие функции f и g, что

$$W \xrightarrow{f} D; \quad (2.2.10)$$

$$P \xrightarrow{g} D \quad (2.2.11)$$

где f – функция, соответствия элементов множества W элементам множества D; g – функция, соответствия элементов множества P элементам множества D.

Таким образом, для для повышения показателей полноты и точности обнаружения DLP-системой угроз информационной безопасности необходимо найти такие функции f и g, чтобы процесс пополнения словарей морфологических описаний слов был максимально упрощен и автоматизирован.

Необходимость разработки быстрого метода идентификации защищаемых данных в передаваемых сообщениях

Для вынесения вердикта v (2.1.1) о возможности дальнейшей передачи сообщения необходимо определить, содержит ли оно хотя-бы один защищаемый факт I. Выявление защищаемых фактов в передаваемом сообщении производится после семантического анализа. Результатом семантического анализа, как правило, является граф, описывающий семантические объекты и связи между ними.

Таким образом, на этапе после семантического анализа в анализаторе DLP-системы имеется два графа – G_1 – граф, описывающий защищаемые данные, и G_2 – граф, описывающий передаваемое сообщение. Тогда для выявления угрозы утечки защищаемой информации необходимо решить задачу поиска изоморфизма графа G_2 в G_1.

По-видимому, первый анализ проблемы изоморфизма графов возникает в статье Р. Рида и Д. Корнейла (1977), с примечательным названием "Graph isomorphism disease". Приведенная в ней библиография из 36 работ и последующая библиография из ещё 32 работ содержат ссылки на большое число алгоритмов, которые по предположению их авторов, распознают изоморфизм произвольных графов за полиномиальное время. Однако, все эти предположения оказались несостоятельными. Наилучший результат к настоящему времени получен Л. Бабаи, Ю. Лаксом и В. Кантором (1983) и опирается на классификацию конечных простых групп.

В решении проблемы изоморфизма графов получены следующие результаты: Изоморфизм n-вершинных графов распознаваем за время $\exp(O(\sqrt{n \log n}))$. Наилучший алгоритм, не использующий теорию групп, был построен М. Годбергом (1983) и имеет сложность $\exp(O(n))$. Упомянем здесь также, что наиболее быстрый с практической точки зрения алгоритм проверки изоморфизма графов принадлежит Б. Маккею и его реализация доступна на его домашней странице [38].

Достаточно быстрый алгортм описан в работе [39]. Его сложность варьируется от $O(n^2)$ в лучшем случае до $O(n! \, n)$ в худшем случае.

	Алгоритм VF2	**Алгоритм Ульмана**
Лучший случай	$O(n^2)$	$O(n^3)$
Худший случай	$O(n! \, n)$	$O(n! \, n^2)$

Табл. 2.2.1. Временная сложность алгоритмов поиска подграфов в графах.

Высокая сложножность алгоритмов определения изоморфизма графов делает их неприменимыми при работе с большими объемами защищаемых данных. Кроме того, необходимо учитывать возможные способы модификации защищаемой информации, что еще больше усложняет анализ.

Таким образом, видится необходимым получить быстрый метод идентификации защищаемых данных в передаваемых сообщениях. Это позволит DLP-системе эффективно работать с большим количеством защищаемых данных.

2.3 Метод снижения числа ошибок первого и второго рода в морфологических анализаторах DLP-систем

Для повышения показателей полноты и точности обнаружения DLP-системой угроз информационной безопасности необходимо повысить эффективность работы функции F_{22} (2.2.5), т.е. повысить качество определения морфологических признаков слов в предложении.

2.3.1 Суть метода

Предлагаемый метод снижения числа ошибок первого и второго рода в морфологических анализаторах DLP-систем основывается на использовании некоторого множества известных, заранее корректных последовательностей морфологических признаков (множество корректных шаблонов) для аналогичных предложений. Аналогичным в данном случае является предложение, имеющее одинаковую длину и последовательность морфологических признаков с анализируемым. Выбор одной из гипотез определяется наличием ее во множестве известных. При этом, если в известном множестве содержится более одной гипотезы для анализируемого предложения, то выбирается наиболее часто встречающаяся.

Важной особенностью метода является то, что необходимое для работы множество корректных шаблонов формируется автоматически на основе специально подготовленного морфологического словаря и не требует предварительно размеченных текстов.

Предлагаемый метод состоит из следующих частей:

1. Формирование множества корректных шаблонов предложений;
2. Формирование множества гипотез о морфологических признаках слов в предложении;
3. Поиск во множестве корректных шаблонов каждой из гипотез;
4. Выделение одной из гипотез на основе результатов поиска.

Формализация задачи представлена следующим образом.

Пусть $T = \{t_i\}$, $i=1,..., n$ – множество корректных шаблонов предложений.

Пусть $H = \{h_i\}$, $i=1,..., m$ – множество гипотез о морфологических признаках слов анализируемого предложения.

Пусть R – правильное описание анализируемого предложения с точки зрения морфологических признаков его слов.

Тогда решаемая задача состоит в том, чтобы выбрать такое $h_j \in H$, что $h_j = R$.

При этом важными подзадачами являются формирование множества корректных шаблонов T и поиск h_i в множестве T.

2.3.2 Формирование множества корректных шаблонов предложений

Сформировать достаточное множество корректных шаблонов вручную не представляется возможным. Если рассматривать 14 частей речи (как один из морфологических признаков) и только предложения длиной до 10 слов, то мощность множества T, необходимого для анализа таких предложений, можно грубо оценить по следующей формуле:

$$|T_{10}| = \sum_{i=1}^{10} 14^i \approx 3 \cdot 10^{11} . \quad (2.3.1)$$

Очевидно, что корректно разметить такое или сравнимое с таким число предложений за разумное время невозможно. Автоматическое решение этой подзадачи «напрямую» сталкивается с исходной решаемой проблемой – с проблемой автоматического определения морфологических признаков в предложениях.

Для решения этой подзадачи была использована следующая идея. Искомое множество T можно формировать на основе предложений, состоящих только из однозначных с точки зрения морфологических признаков слов.

Таким образом, процесс формирования множества корректных шаблонов предложений состоит из следующих шагов.

1. Формирование морфологического словаря, в котором исключены все омонимы.
2. Определение предложений, состоящих только из полученных на шаге 1 слов.
3. Создание шаблонов на основе полученных на шаге 2 предложений.
4. Добавление полученных шаблонов в искомое множество.

На шаге 1 для каждой словоформы в словаре ищется совпадающая с ней, но отличающаяся морфологическим описанием. Если найдены совпадающие словоформы, у которых отличается морфологические признаки, то эта

словоформа не добавляется в словарь. Те в создаваемый словарь попадают словоформы, которые имеют единственное морфологическое описание.

Шаги 2–4 являются сугубо техническими и не представляют интереса.

Описанный способ дает возможность получить необходимое множество T автоматически, без использования предварительно размеченных текстов. Таким образом, реализуется первая часть предлагаемого метода.

2.3.3 Формирование множества гипотез о морфологических признаках слов в предложении

Следующей подзадачей является формирование множества гипотез H о морфологических признаках слов в предложении. Для этого для каждого слова в предложении из словаря извлекается список возможных морфологических признаков. Далее, с помощью перестановки возможных морфологических признаков для каждого слова формируется искомое множество H.

2.3.4 Поиск в множестве корректных шаблонов каждой из гипотез

Полученное в первой части множество корректных шаблонов S предложений достаточно велико (2.3.1). Поиск перебором в таком множестве будет крайне неэффективен. В связи с этим предлагается организовать хранение множества корректных шаблонов с помощью словаря.

В нашей реализации описываемого метода шаблон предложения s_i представляет собой последовательность чисел, т.е. морфологические характеристики кодируются числами.

$t_i = \{m_i\}$, $i=1,..., n$, m_i – закодированная морфологическая характеристика i-го слова в предложении.

В качестве примера можно рассмотреть шаблон, полученный из предложения «*Документы были отправлены заказчику утром*.». Для наглядности рассмотрим только один морфологический признак – часть речи. В нашем случае кодирование выполнялось следующим образом (Таблица 2.3.1).

Часть речи	**Код**
Существительное	0x0001

Глагол	0x0002
...	
Наречие	0x0100
Предлог	0x0200

Таблица 2.3.1. Кодирование частей речи числовыми значениями

Таким образом, шаблон, полученный из указанного выше предложения, имеет вид:

$$t = \{1, 2, 2, 1, 256\}.$$

Для организации хранения и поиска по большому числу таких числовых последовательностей был выбран словарь неограниченной вложенности, где на каждом уровне ключем является код части речи, а значением – кортеж из словаря следующего уровня и числа, означающего количество встреч предложения с частями речи, которые были закодированны использованными числами.

$T[m_1]\ [m_2]\ldots\ [m_n] = (T_{n+1},\ N)$, где T_{n+1} – словарь следующего уровня вложенности, а N – число встреч шаблона t_i.

Такая структура хранения довольно проста в реализации и использовании, и позволяет выполнять поиск по большому числу шаблонов предложений с приемлемой скоростью, что подтверждается экспериментом.

2.3.5 Выделение одной из гипотез на основе результатов поиска

Поиск каждой из гипотез h_i в множестве известных шаблонов T может привести к трем различным исходам:

1. $H \cap T = \varnothing$ – в имеющемся множестве корректных шаблонов не найдено ни одной гипотезы;
2. $H \cap T = \{ h_j \}$ – в имеющемся множестве корректных шаблонов найдена одна гипотеза;
3. $H \cap T = \{ h_j \ldots h_k \}$ – в имеющемся множестве корректных шаблонов найдено несколько гипотез.

Большое число результатов первого исхода говорит о том, что имеющегося множества корректных шаблонов недостаточно, и требуется его пополнение.

В случае второго исхода единственная найденная гипотеза h_j считается верной.

В случае третьего исхода выбирается гипотеза, которая встречалась наиболее часто.

Полученная в результате гипотеза h_j является в среднем более корректной, чем случайно выбранная из исходного множества H. Это подтверждается экспериментом.

Таким образом, предложенный метод позволяет повысить качество определения морфологических характеристик слов в предложении, что решает поставленную задачу повышения показателей полноты и точности обнаружения DLP-системой угроз информационной безопасности.

2.4 Метод предотвращения передачи конфиденциальных ЕЯ собщений, содержащих отсутствующие в словаре парадигмы слов

Для повышения показателей полноты и точности обнаружения DLP-системой угроз информационной безопасности необходимо найти такие функции f и g (2.2.10), (2.2.11), чтобы процесс пополнения словарей морфологических описаний слов неизвестными словоформами был максимально упрощен и автоматизирован.

Основой предлагаемого предметно-ориентированного морфологического анализатора, содержащего идентификационные признаки словоформ предметной области, разработанного для русского языка, служит словарь А.А. Зализняка [28].

Пусть $W = \{ w_i \}$, $i=1,..., n$ – множество исходных форм слов в БД DLP-системы.

Пусть $P = \{ p_j \}$, $j=1,..., k$ – множество парадигм исходных форм слов.

Каждому элементу множеств P и W соотвествует морфологическое описание D_j: $p_j \rightarrow D_j$, $w_j \rightarrow D_j$.

Пусть v – словоформа.

Пусть $E = \{ e_r \}$, $r=1,..., z$ – множество стандартных окончаний слов.

Тогда необходимо найти такие функции f и g, что

$$W \xrightarrow{f} D \quad (2.2.10)$$

$$P \xrightarrow{g} D \quad (2.2.11)$$

где f – функция, соответствия элементов множества W элементам множества D; g – функция, соответствия элементов множества P элементам множества D.

Предлагаемый метод решения основан на том, что любой словоформе v сопоставим класс основ B и класс окончаний (флексий) E, из которого состоит данная словоформа.

$$\forall v \longrightarrow \{ B; E \} \quad (2.2.12)$$

Для каждого слова W БД СЗИ можно выделить морфологический класс k его парадигм P_k, такой, что словоформа данного морфологического класса (v_k) входит в множество парадигм этого класса и выражается суммой основ и окончаний слова данного морфологического класса.

$$v_k \in P_k = B_k + E_k \quad (2.2.13)$$

Соответствие $P_j \rightarrow D_j$, позволяет получить морфологический и идентификационный признак, содержащий информацию, используемую для обнаружения угроз информационной безопасности (морфологический шаблон).

Рассмотрим существующий морфологический словарь $Z = \{ z_i \}$, каждая запись z_i в котором имеет структуру

$$z_i = \{ v_i; W_i; D_i \}, \quad (2.2.14)$$

т.е. состоит из словоформы v_i, исходной формы слова W_i и морфологического описания D_i.

Словарь Z, на сегодняшний день, содержит более 2,5 млн словоформ. Задача состоит в том, чтобы учитывая регулярность русского языка [29], анализировать сообщения ЕЯ, которые содержат отутсвующие в словаре термины, а также поддерживать актуальность и полноту словарной базы данных с наименьшими трудозатратами. Решение поставленной задачи основывается на словаре, содержащем морфологические описания словоформ А.А. Зализняка, содержащем

только базовые словоформы русского языка и множество соответствующих им окончаний.

Рассмотрим как образованы словоформы слов «ПРЕОБРАЗОВАТЕЛЬ» и «СЧИТЫВАТЕЛЬ» (Таблица 2.4.2).

Словоформы ($W, P)_j$		**Морфологические описания D_j**
ПРЕОБРАЗОВАТЕЛЬ	СЧИТЫВАТЕЛЬ	Сущв Муж Неодуш Им, Вин
ПРЕОБРАЗОВАТЕЛЯ	СЧИТЫВАТЕЛЯ	Сущв Муж Неодуш Род
ПРЕОБРАЗОВАТЕЛЮ	СЧИТЫВАТЕЛЮ	Сущв Муж Неодуш Дат
ПРЕОБРАЗОВАТЕЛЕМ	СЧИТЫВАТЕЛЕМ	Сущв Муж Неодуш Тв
ПРЕОБРАЗОВАТЕЛЕ	СЧИТЫВАТЕЛЕ	Сущв Муж Неодуш Пред
ПРЕОБРАЗОВАТЕЛИ	СЧИТЫВАТЕЛИ	Сущв Муж Неодуш Им, Вин
ПРЕОБРАЗОВАТЕЛЕЙ	СЧИТЫВАТЕЛЕЙ	Сущв Муж Неодуш Род
ПРЕОБРАЗОВАТЕЛЯМ	СЧИТЫВАТЕЛЯМ	Сущв Муж Неодуш Дат
ПРЕОБРАЗОВАТЕЛЯМИ	СЧИТЫВАТЕЛЯМИ	Сущв Муж Неодуш Тв
ПРЕОБРАЗОВАТЕЛЯХ	СЧИТЫВАТЕЛЯХ	Сущв Муж Неодуш Пред

Таблица 2.4.2. Словоформы и их морфологический описания

Из таблицы видно, что флексия словоформ одинакова, т.е. они получены из базовой формы *W* одинаковым образом, путем добавления соответсвующих окончаний *E*. Следовательно, достаточно иметь морфологические описания *D* словоформ слова «ПРЕОБРАЗОВАТЕЛЬ», чтобы построить аналогичные описания для словоформ слова «СЧИТЫВАТЕЛЬ».

На основе этой идеи разработан предлагаемый метод анализа сообщений ЕЯ, содержащих отсутствующие в словаре парадигмы слов. Он состоит из следующих частей:

1. Разбор словаря Зализняка, генерация всех словоформ на основе исходных форм слова;
2. Разбор словаря с некоторыми морфологическими описаниями вида, который описан выше;
3. Сопоставление словоформ из словарей, полученных на первых двух шагах с целью выделения характерных морфологических описаний для каждого

окончания в рамках класса слова (класса его флексий), к которому они относятся;

4. Определение класса слов из передаваемого сообщения, отсутствующих в морфологическом словаре;
5. На основе множества соответствий вида «класс слова, окончание» – «морфологическое описание», полученных на третьем шаге, словоформам из передаваемого сообщения, отсутствующим в морфологическом словаре, дается морфологическое описание.

Задачи на первых двух шагах являются чисто техническими, и их описание не представляет какого-либо интереса.

Выделение характерных морфологических описаний для каждого окончания, описанное на шаге 3, осуществляется следующим образом. Каждое окончание входит в свой «класс» окончаний. Для слова «ПРЕОБРАЗОВАТЕЛЬ» это окончания «Я», «Ю», «ЕМ», «Е», «И», «ЕЙ», «ЯМ», «ЯМИ» и «ЯХ». Окончание «Ю» слова «ЗЕМЛЮ» хотя и совпадает с окончанием «Ю» слова «ПРЕОБРАЗОВАТЕЛЮ», но входит в совершенно другой класс, и поэтому будет иметь другой набор морфологических описаний. Кроме класса также учитывается часть речи слова и одушевленность/неодушевленность для имен существительных.

Таким образом, полный ключ в ассоциативном массиве с морфологическими описаниями состоит из «класса» окончания, части речи и признака одушевленности/неодушевленности для имен существительных. Таким образом, в случае, когда одна исходная форма относится к разным частям речи, для каждой части будет храниться свой набор морфологических описателей.

Определение класса слов из передаваемого сообщения, отсутствующих в морфологическом словаре, может быть реализовано методами, описанными в работах [27], [30] и [60].

Полученые на третьем шаге соответствия применяются на пятом шаге для составления словаря с морфологическими описаниями. Из сгенерированного на основе словаря Зализняка списка словоформ берется словоформа, а затем по

ключу, описанному выше, в ассоциативном массиве находится морфологическое описание для этой словоформы. Таким образом, реализуются функции f и g, приведенная в формулах (2.2.10), (2.2.11) [31].

За счет того, что русский язык достаточно регулярен, и многие слова формируются похожим способом, появляется возможность автоматически получить для исходной формы слова морфологические описание всех его словоформ.

Метод предотвращения передачи конфиденциальных ЕЯ собщений, содержащих отсутствующие в словаре парадигмы слов, позволяет автоматически получить морфологическое описание несловарного термина в анализируемом сообщении и пополнить морфологический словарь всеми его словоформими. Благодаря этому DLP-система может более корректно анализировать характерные для современных ИС ЕЯ-сообщения. Также появляется возможность уйти от последовательного внесения в морфологический словарь всех возможных словоформ с их морфологическими характеристиками, что является необходимой, но нетипичной задачей для служб ИБ и ИТ.

2.5 Метод идентификации защищаемых данных в передаваемых сообщениях на основе анализа связей в объектной модели естественного языка

2.5.1 Суть предлагаемого метода

Как было показано в разделе 2.2, обнаружение защищаемых данных (фактов *I*) с использованием графов, которые описывают семантические связи между анализруемыми объектами, является затруднительным и, во многих случаях, неприменимым на практике.

Суть анализа графов, описывающих семантические связи между объектами и их свойствами, состоит в том, что происходит сравнение объектов и их характеристик. Поскольку реализация на графах достаточно точно описывает объектную модель, можно сделать вывод, что любая система, основанная на сравнении семантических объектов будет иметь сравнимую производительность. Поэтому исследование в этом направление не видится перспективным. При этом

необходимо отметить, что такие способы применимы при небольших объемах защищаемых данных.

Основной идеей предлагаемого метода выявления защищаемых данных является использование для сравнения связей объектов, вместо самих объектов.

Рассмотрим предложение «Планируется размещение в датацентре в ближайшее время.» Предположим, что эта информация является коммерческой тайной или ее частью, и она не должна покидать пределы ИС компании, которая является ее владельцем. Семантические связи в этом предложении можно видеть на рис. 2.5.1.

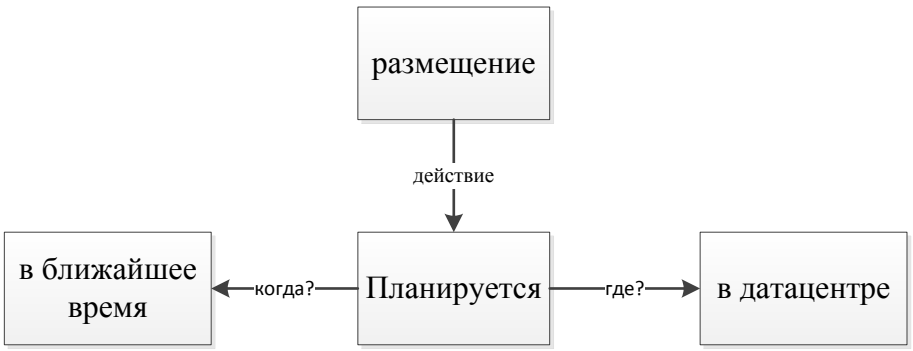

Рис. 2.5.1 Семантические связи в «защищаемом» предложении

Вместо сравнения объекта («размещение») и его характеристик («планируется» и т.д.) предлагается сравнивать связи между объектами: «действие» между «размещение» и «планируется», «когда?» между «планируется» и «в ближайшее время» и «где?» между «планируется» и «в датацентре». Таким образом, в результате анализа защищаемых данных синтаксический анализатор DLP-системы будет обладать набором связей с определенными характеристиками.

При передаче сообщения, подлежащего анализу, синтаксический анализатор DLP-системы будет строить такой-же набор связей для передаваемого сообщения. Предположим, что передается сообщение «Рассчитываем вскоре разместиться в DC.» Рассмотрим семантические связи, которые будут получены в результате анализа этого сообщения (рис. 2.5.2).

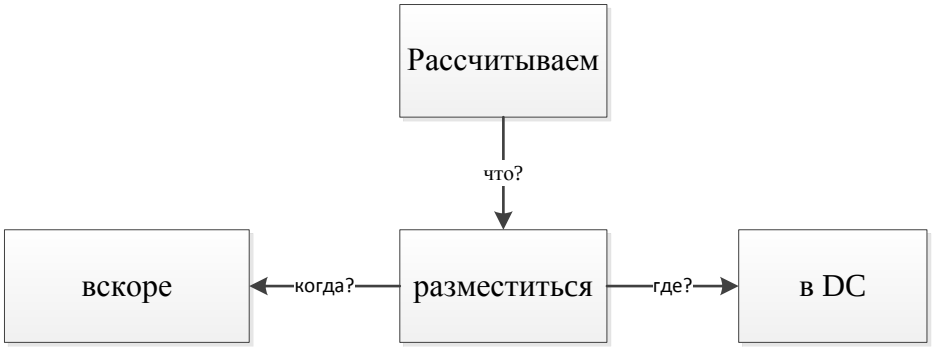

Рис. 2.5.2 Семантические связи в передаваемом предложении

После несложного преобразования, которое может быть выполнено автоматически, можно получить семантические связи на рис. 2.5.3.

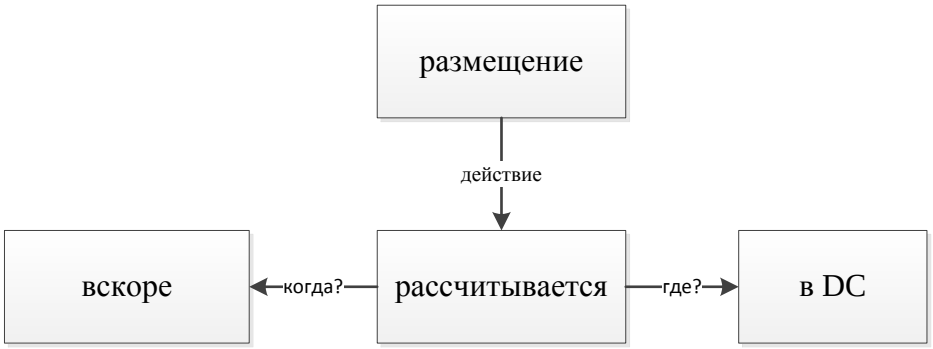

Рис. 2.5.3 Семантические связи в передаваемом предложении после одного из преобразований

Таким образом, в результате анализа передаваемого сообщения может быть получен следующий набор семантических связей: «действие» между «размещение» и «рассчитывается», «когда?» между «рассчитывается» и «вскоре» и «где?» между «рассчитывается» и «в DC».

Сравнивая рис. 2.5.1 и 2.5.3, а также наборы семантических связей, полученных в результате анализа защищаемого и передаваемого текстов можно отметить их похожесть. Суть предлагаемого метода состоит в том, чтобы автоматически строить множества таких связей и сравнивать их. При обнаружении похожести можно сделать вывод, что передаваемое сообщение содержит защищаемые факты *I*.

Как будет показано в дальнейшем, выполнение таких операций может быть организовано достаточно быстро, что решает проблему производительности с большими объемами данных при использовании классического метода сравнения объектов.

2.5.2 Формальное описание метода

Рассмотрим множество *D* защищаемых данных и множество *I* защищаемых фактов (п. 2.1)

Пусть $L = \{ l_i \}$, $i=1,..., n$ – множество связей, полученных в результате анализа защищаемых данных *D*.

Пусть $L_t = \{ l_i \}$, $i=1,..., n$ – множество связей, полученных в результате анализа сообщений, передаваемых по одному из разрешенных каналов связи *C*.

Пусть функция $F_{match}(L, L_t)$ – функция выявления похожести множества связей L_t с одним из подмножеств множества связей L.

$$F_{match}(L, L_t) = \text{c}, \quad (2.5.2.1)$$

где c – мера похожести множеств связей L_t и L.

Если $\text{c} \geq c_{L_t}$, где c_{L_t} – некая константа для данного L_t, то считается что множество связей L_t похоже на одно из подмножеств множества связей L. Применительно к анализу угрозы утечки защищаемой информации это говорит о наличии в передаваемом сообщении одного из защищаемых файтов *I*.

Предлагаемый метод идентификации защищаемых данных в передаваемых сообщениях состоит из следующих этапов:

1. Формирование множества связей L защищаемых данных в результате анализа множества защищаемых данных *D*
2. Формирование множества связей L_t передаваемых сообщений в результате анализа ЕЯ-текстов, передаваемых по одному из каналов *C*
3. Вычисление функция $F_{match}(L, L_t)$ и константы c_{L_t}, сравнение результатов

2.5.3 Существенная особенность анализа естественноязыковых сообщений с целью выявления угрозы утечки информации

Важной особенностью и отличием задачи выявления угрозы утечки информации от других задач анализа естественного языка (например, поисковых задач) является то, что на каждом уровне анализа мы можем и должны учитывать все возможные гипотезы. Например, при невозможности разрешить неоднозначность в определении части речи или других морфологических характеристик слова, на следующий уровень анализа должны передаваться все возможные гипотезы.

Таким образом, на уровне синтаксического анализа с большой вероятностью будет иметься множество различных гипотез описания передаваемого сообщеия и защищаемых данных.

При анализе естественноязыковых сообщений с целью выявления угроз ИБ нет ничего плохого в том, что будут обнаружены неестественные или неочевидные трактовки передаваемого текста: если передаваемое сообщение можно тракотовать как защищаемый текст, то не важно, что именно имел ввиду отправитель. Система защиты от утечек должна безусловно блокировать передачу такого сообщения.

Таким образом, неверная (точнее, непредполагаемая), но формально допустимая интерпретация передаваемого сообщения не может вызвать ошибку при выявлении утечек информации – «ложное» срабатывание в данном случае не является ложным.

2.5.4 Формирование множеств семантических связей

В связи с особеностями анализа естественноязыковых сообщений, описанных в предыдущем пункте, после этапа семантического анализа рассматриваются все полученные гипотезы. В дальнейшем в работе рассматривается одна отдельная гипотеза, и считается очевидным, что все описанные шаги должны быть применены к каждой из полученных гипотез.

При формировании множества семантических связей необходимо учитывать, что для вычисления функции похожести необходимо, чтобы каждая связь была представлена числом, и в то-же время, удовлетворяла следующим условиям:

1. Связи различных типов не должны пересекаться, какие бы термины они не связывали
2. Связи близких терминов должны иметь близкое значение

Для реализации требования 1 достаточно для каждого типа связи разбить числовую ось от нуля до L_MAX на N отрезков, где L_MAX – максимальное значение, которое может быть сопоставлено связи, а N – число типов связи. Таким образом, в диапазон значений

$$((k-1)\cdot\frac{L_{MAX}}{N}, k\cdot\frac{L\_MAX}{N})$$

попадают вся связи *k*-ого типа, $k \in [1, N]$.

Для реализации требования 2, смысл которого подробно описан в п. 2.5.5, предлагается следующий подход:

Каждому термину, который может быть связан связью типа *k*, ставится в соответствие число *T*. Организуем соответствие так, чтобы для двух близких терминов (синонимов) числа *T* были близки. Для этого можно, например, воспользоваться словарем синонимов.

В качестве простейшего способа назначения T_i можно предложить положительные числа с шагом s = 11, так чтобы к каждому слову можно было приписать 9 синонимов, заняв, таким образом, оставшиеся до следующего интервала значения.

Рассмотрим пример назначения T_i для слова «выполнить». Пусть слову «выполнить» будет поставлено в соответствие число $T_i = 100$. Тогда для синонимы слова «выполнить» будут иметь следующие значения:

T_i	Термин
100	слушаться
101	соблюсти
102	претворить

103	осуществить
104	исполнить
105	**выполнить**
106	сделать
107	реализовать
108	воплотить
109	провести
110	удовлетворить

Таблица 2.5.4.1 Соответствие параметра T_i терминам

В результате, каждому термину для каждого типа связи k поставлено в соответствие несколько чисел T_i:

- Основное значение T_i. Одно значение числа T_i каждый термин получает «по-определению», это значение будет характеризовать его точное употребление. Например, для слова «выполнить» из примера выше это будет значение $T_i = 105$.
- Дополнительные значения T_i. Несколько других значений, в зависимости от числа вхождений этого термина в списки синонимов для других слов. Так, например, для слова «исполнить» из примера выше значение $T_i = 104$ будет дополнительным, поскольку в данном случае оно входит в список как синоним.

Таким образом, каждой связи двух терминов L_i типа k ставится в соответствие одно или несколько чисел T_{ik}, таких что:

$$T_{ik} = (k-1) \cdot \frac{L\_MAX}{N} + T_{i1} + T_{i2}, \quad (2.5.4.1)$$

где T_{i1} – значение параметра T_i для первого термина связи, T_{i2} – значение параметра T_i для второго термина связи. Такой выбор значения для T_{ik} позволяет для каждой связи находить близкие по «смыслу» связи, что позволяет существенно расширить возможности определения угрозы утечки конфиденциальной информации.

2.5.5 Вычисление функции определения похожести множеств связей

На последнем этапе предлагаемого метода необходимо вычислить функцию определения похожести $F_{match}(L, L_t)$ множеств связей L и L_t.

Напомним, что если $F_{match}(L, L_t) \geq c_{L_t}$, где c_{L_t} – некая константа для данного L_t, то считается что множество связей L_t похоже на одно из подмножеств множества связей L. Применительно к анализу угрозы утечки защищаемой информации это говорит о наличии в передаваемом сообщении одного из защищаемых файтов I.

Для вычисления функции $F_{match}(L, L_t)$ воспользуемся аппаратом из области цифрового спектрального анализа. Для выявления сходства двух сигналов используется взаимокорреляционная функция (ВКФ). Для двух дискретных сигналов u и v ВКФ может быть определена по формуле:

$$B(n) = \sum_{j=-\infty}^{+\infty} u_j v_{j-n}, \quad (2.5.5.1)$$

где n – целое число, положительное, отрицательное, или нуль. [36]

Выбор ВКФ для рассчета функций похожести $F_{match}(L, L_t)$ множеств связей L и L_t сделан не случайно. Во-первых, эта функция решает поставленную задачу – определеяет «похожесть» функций. Во-вторых, такой способ является быстрым (в сравнении с задачей определения изоморфизма графов для «классического» подхода семантического анализа) за счет возможности использования метода быстрого преобразования Фурье (БПФ).

Надо отметить, что множества связей L и L_t в том виде, как они описаны в п. 2.5.4 для рассчета функций похожести $F_{match}(L, L_t)$ не подходят.

Во-первых, ВКФ оперирует совокупностями отсчетов, следующих во времени с одинаковым интервалом. Для этого введем последовательность $L' \ni L$, дополняющее L нулевыми отсчетами с интервалом 1.

Во-вторых, ВКФ исходных последовательностей числел из L' и L'_t будет в большинстве случаев равняться нулю или не превышать некоторого небольшого значения, что не позволит сделать какие-то выводы. Рассмотрим пример, когда L= $\{105, 155, 200, 220\}$

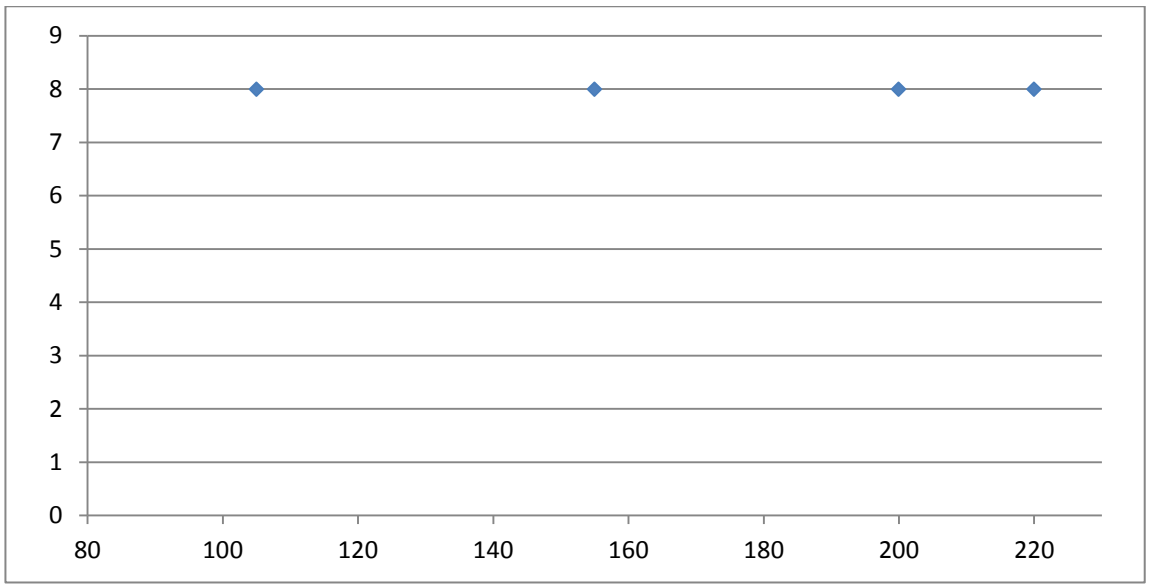

Рис. 2.5.5.1 Последовательность L и соответствующие ее элементам веса

Для того, чтобы учесть возможность замены некоторых терминов синонимами, воспользуемся тем, что связи синонимов имеют близкие значения. Для обеспечения этого свойства связей было выдвинуто ребование 2 п. 2.5.4: связи близких терминов должны иметь близкое значение.

Существует два способа учета синонимов:

1. Заведомо расширить множество L значениями T_i синонимов каждого слова. В таком случае, если у каждого слова из связи есть n синонимов, то для каждой связи прийдется вычислять n^2 значений T_{ik} и хранить их в памяти.
2. При вычислении функции корреляции перебрать значения S_{MAX}^2 значений $B(n)$ и выбрать максимальное.

В первом случае функцию корреляции $B(n)$ (2.5.5.1) можно вычислить только для $n = 0$, однако для хранения синонимов потребуется больше памяти.

Во втором случае памяти потребуется существенно меньше, однако при анализе каждого сообщения потребуется совершить существенно больше вычислений.

В дальнейшем будет рассматриваться более быстрое решение, на основе способа 1.

Воспользуемся тем, что связи синонимов имеют близкие значения для расширения множества L. В результате, получим следующее множество L.

$$L = \{102,103,104,105,106,107,108, \dots ,217,218,219,220,221,222,223\}.$$

Необходимо обратить внимание на то, что дополнять синонимами необходимо только последовательность L. Последовательность L_t в дополнении не нуждается.

Также надо учитывать, что замена синонимом не является полноценной. Поэтому будем уменьшать «вес» замены по мере удаления от исходного слова.

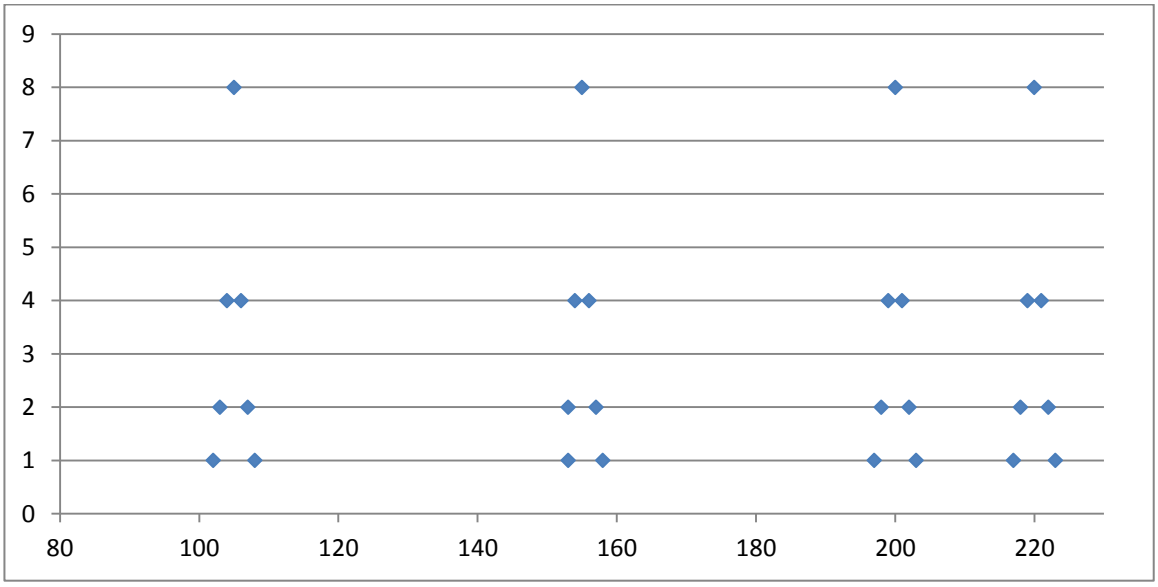

Рис. 2.5.5.2 Дополненная последовательность L и соответствующие ее элементам веса

После формирования множеств L' и L'_t можно приступать к вычислению $F_{match}(L, L_t)$.

Предлагается следующий способ вычисления $F_{match}(L, L_t)$.

1. При разном количестве отсчетов L' и L'_t увеличиваем длину более короткой последовательности, дополняя ее $M - 1$ нулевыми отсчетами. Как правило, это L'_t
2. Далее вычисляется сумма

$$F_{match}(L, L_t) = \sum_{i=0}^{N} L'_i \cdot L'_{ti}, \quad (2.5.5.2)$$

где $N = |L'_t|$ после возможного дополнения нулями на шаге 1, L'_i – значение i-ого элемента последовательности.

3. Далее вычисляется константа c_{L_t} (2.5.2.1) для сравнения с $F_{match}(L, L_t)$

$$c_{L_t} = 2 \cdot |L_t| \text{ (2.5.5.3)}$$

Вычисление константы c_{L_t} указанным в (2.5.5.3) способом не всегда корректно. Описанный способ подходит для случаев, когда передаваемое сообщение содержит только защищаемые данные. Доработка метода для случаев, когда передаваемое сообщение содержит не только защищаемые данные является одним из направлений для дальнейшего развития метода.

Предложенный метод идентификации защищаемых данных в передаваемых сообщениях позволяет обнаружить угрозу утечки информации как при передаче защищаемого текста без модификации, так и в случае замены ряда терминов и иной формулировки защищаемых данных. Важной особенностью является то, что предложенный метод работает быстрее, чем методы основанные на графах семантических связей, что позволяет использовать его при работе с большими объемами защищаемых данных.

2.6 Выводы

1. На основе формализованного описания современной защищаемой информационной системы разработана модель угрозы утечки конфиденциальной информации.

2. Приведено общее описание DLP-системы, на основании которого разработана формальная модель работы DLP-системы.

3. Анализ полученной модели угрозы утечки конфиденциальной информации в сочетании с формальной моделью работы DLP-системы показал возможные направления развития существующих методов анализа естественноязыковых сообщений в DLP-системах.

4. В рамках выбранных направлений развития морфологических анализаторов DLP-систем был разработан метод снижения числа ошибок первого и второго рода в морфологических анализаторах DLP-систем.

5. В рамках выбранных направлений развития морфологических анализаторов DLP-систем был разработан метод метод предотвращения передачи конфиденциальных ЕЯ собщений, содержащих отсутствующие в словаре парадигмы слов.

6. В рамках выбранных направлений развития морфологических анализаторов DLP-систем был разработан метод идентификации защищаемых данных в передаваемых сообщениях на основе анализа связей в объектной модели естественного языка

Применение разработанных методов позволяет решить поставленную задачу повышения показателей качества защиты DLP-систем.

3. Сравнительный анализ

3.1 Оценка показателей качества предложенных решений

Оценка показателей качества метода снижения числа ошибок первого и второго рода в морфологических анализаторах DLP-систем

Описанный в работе метод снижения числа ошибок первого и второго рода в морфологических анализаторах DLP-систем был реализован и встроен в подсистему определения морфологических характеристик слов в предложении.

Показатели качества разработаного метода оценивались сравнением результатов работы подсистемы с заведомо корректными на размеченных вручную текстах.

В качестве первого источника размеченных текстов изначально был выбран Национальный корпус русского языка [32]. Но он не подошел из-за большого числа использованных там «композитных» частей речи, например «местоимение–существительное», «местоимение–прилагательное», «числительное–прилагательное» и т.д. Поскольку в используемом нами словаре такие «части речи» не используются, статистика получалась искаженная, и для сравнения был выбран Открытый корпус русского языка [33].

Сравнение производилось на случайной выборке размеченных вручную текстов Открытого корпуса русского языка. В ходе каждого опыта сравнивалось 17130 слов в 2300 предложениях.

На рис. 3.1.1 показана зависимость процента исправленных ошибок от количества использованных при анализе шаблонов предложений.

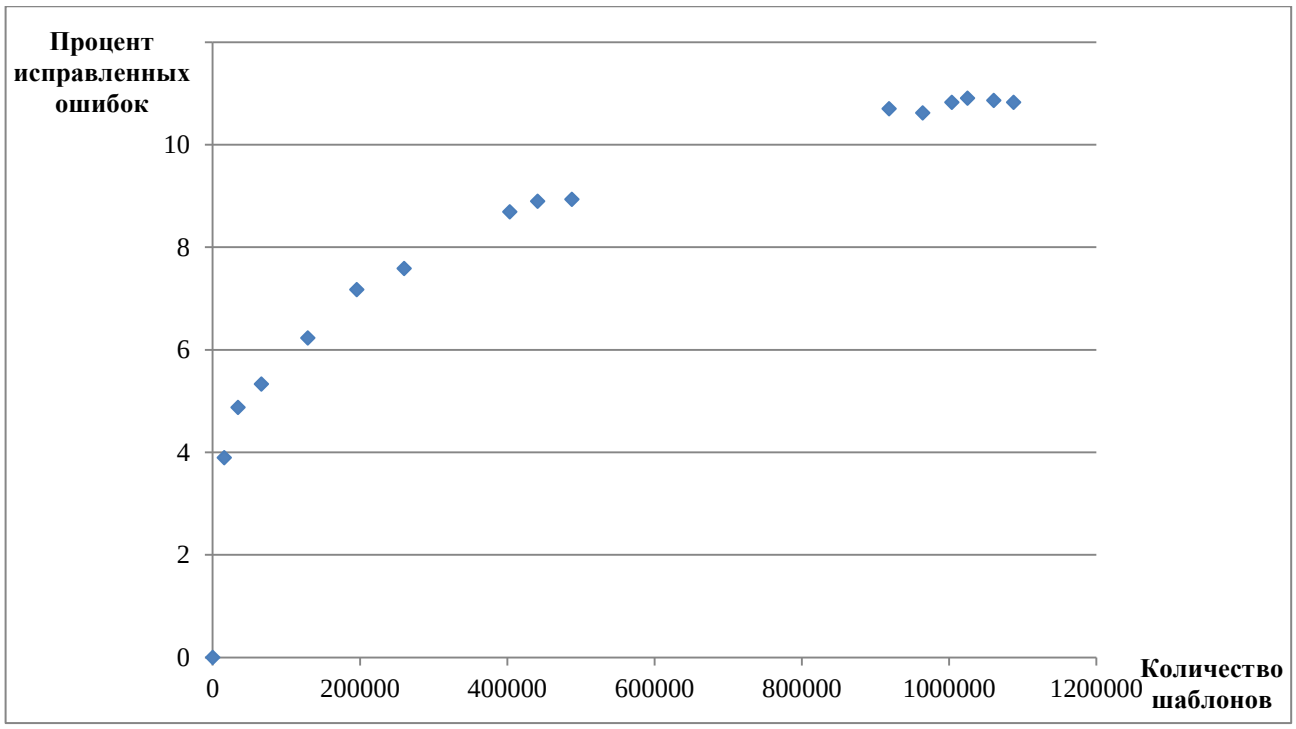

Рис. 3.1.1. Зависимость процента исправленных ошибок определения морфологических характеристик слов от числа шаблонов предложений, использованных при анализе

При этом имеющееся множество шаблонов далеко от насыщения. На рис. 3.1.2 показана зависимость количества полученных шаблонов от числа разобранных текстов.

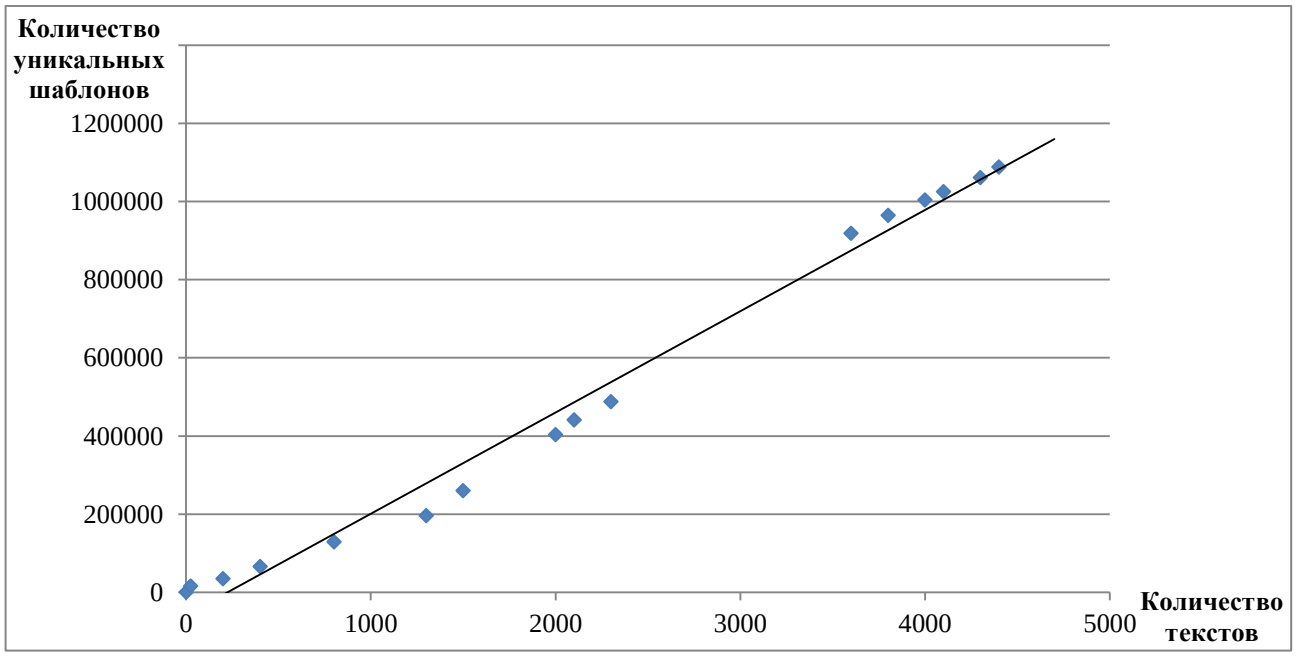

Рис. 3.1.2. Зависимость количества полученных шаблонов от числа разобранных текстов

По графику видно, что число шаблонов линейно возрастает в зависимости от числа разобранных текстов, что говорит о возможности улучшения полученных нами параметров. По приблизительной оценке, при мощности множества шаблонов ~10^9, можно снижения числа ошибок разбора на 30-35%.

Полученная на выходе системы гипотеза является в среднем более корректной, чем случайно выбранная из исходных. Иными словами, число верных совпадений морфологических характеристик слов в предложении выше, чем у случайно выбранной гипотезы, что подтверждено экспериментом.

Оценка показателей качества метода предотвращения передачи конфиденциальных ЕЯ собщений, содержащих отсутствующие в словаре парадигмы слов

Для оценки показателей качества метода предотвращения передачи конфиденциальных ЕЯ собщений, содержащих отсутствующие в словаре парадигмы слов, обозначим количество правильных извлечений системы анализа DLP-фильтра h, количество требуемых извлечений d, а общее количество извлечений n. Тогда для полноты R и точности P выявления угроз утечки в передаваемом сообщении справедливы следующие соотношения:

$$R_i = \frac{h_i}{d_i} \text{ и } P_i = \frac{h_i}{n_i}.$$

Эксперимент по поиску с использованием словарей проводился на основе случайной выборки предложений из национального корпуса русского языка [32]. Объем выборки – 180 тыс. словоупотреблений, из которых 90 тыс. – пресса и по 30 тыс. из научных текстов, художественных текстов и законодательства.

Для проведения эксперимента была разработана простая поисковая система, использующая в своей основе булевскую модель поиска [35]. Разработанная система позволяла автоматически формировать поисковые запросы и

обрабатывать результаты поиска. Таким образом, значение d числа требуемых извлечений было известно при формировании поисковых запросов, что обеспечивало правильность полученного результата. Общее количество извлечений p и количество правильных извлечений h вычислялись в ходе эксперимента, после обработки каждого поискового запроса.

В первом случае, поисковая система использовала словарь Зализняка и словарь с полными морфологическими описаниями для только одного слова каждого класса. Во втором случае, использовался словарь, который сгенерирован с помощью описанного выше метода.

Необходимо отметить, что вместо реализации шага 4 предложенного метода (определение класса слов из передаваемого сообщения, отсутствующих в морфологическом словаре) класс слова определялся по словарю Зализняка. Т.е. брались те слова, которые отсутствуют в словаре морфологических описаний слов Z, но присутствуют в словаре Зализняка.

В ходе эксперимента измерялись полнота (R) и точность (P) поиска на случайной выборке из национального корпуса русского языка. Результаты измерения приведены на графике на рис. 1, 2.

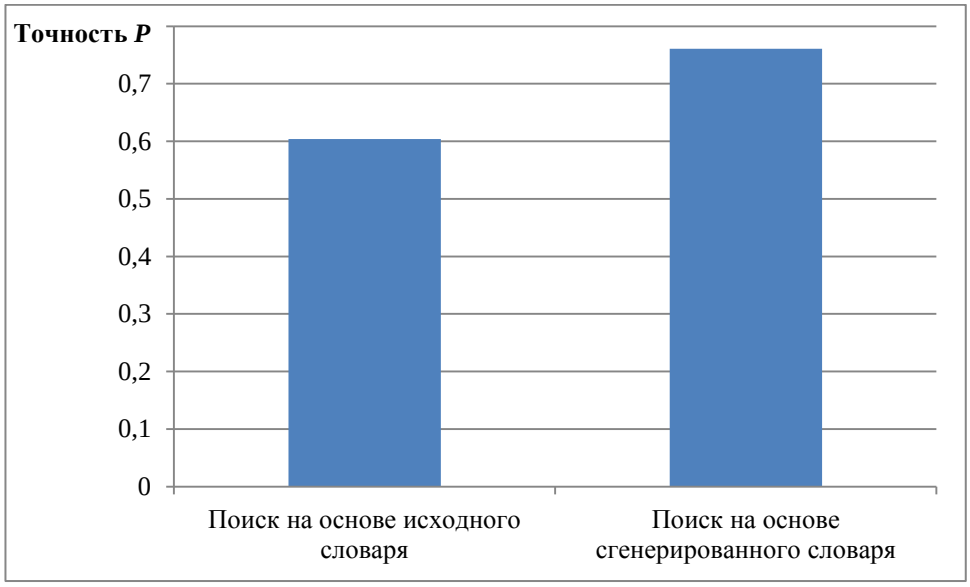

Рис. 3.1.3. Результаты измерения точности P

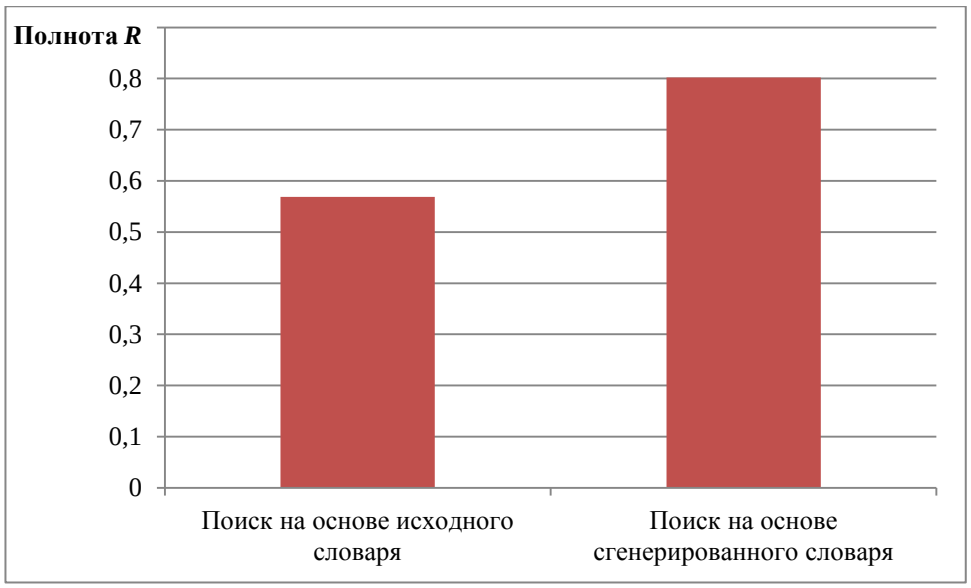

Рис. 3.1.4. Результаты измерения полноты *R*

Эксперимент показал, что при использовании сгенерированного описанным выше методом словаря точность поиска возрола на 20%, а полнота на 29%. Следовательно, использование описанного метода может увеличить вероятность корректного распознавания естественноязыковых конструкций морфологическим анализатором DLP-системы, что решает поставленную задачу повышения показателей качества фильтрации DLP-систем.

Оценка показателей качества метода идентификации защищаемых данных в передаваемых сообщениях на основе анализа связей в объектной модели естественного языка

Предложенный метод выявления защищаемых данных на основе анализа связей в объектной модели естественного языка, в отличие от предыдущих двух методов, не проверен экспериментально. Это связано с тем, что реализация проверки показателей качества выявления одних текстов естественного языка в других крайне затруднительна. Среди основных проблем, возникающих при попытке сравнить показатели качества решения этой задачи, можно отметить следующие:

1. Высокая сложность качественной реализации некоторых этапов морфологического анализа (2.2.5)

2. Отсутствие подготовленных данных (защищаемых текстов ЕЯ и текстов ЕЯ, содержащих в себе различными способами измененные факты защищаемых текстов), на которых можно проводить эксперименты по оценке эффективности тех или иных методов
3. Отсутствие открытых реализаций других методов идентификации защищаемых данных в передаваемых сообщениях, которые необходимы для сравнения

Последняя проблема связана со спецификой разработки программых и аппаратных продуктов в области ИБ, а также узостью и специфичностью задач, которые решаются DLP-системами.

В связи с этим в настоящий момент возможно только теоретическое сравнение производительности предложенного метода.

Для построения семантической модели текстов естественного языка сейчас, как правило, используются графы. В частности – деревья. [40][41][57][58]. В таком случае основная задача DLP-системы – задача поиска защищаемого факта в передаваемом сообщении – сводится к задаче поиска изоморфизма двух графов.

Рассмотрим сначала теоретические оценки сложности решения этой задачи.

Пусть V – конечное множество и E – подмножество множества его двухэлементных подмножеств. Тогда пара $G = (V, E)$ называется (простым) графом с множеством вершин V и множеством ребер E. Будем говорить, что вершины u и v графа G смежны, если последний содержит ребро $(u, v) := \{u, v\}$.

Граф $G = (V, E)$ может быть задан матрицей смежности $A = A(G)$ графа G, т.е. квадратной $n \times n$ матрицы, определяемой следующим образом:

$$A_{uv} = \begin{cases} 1, \text{если } (u, v) \in E \\ 0, \text{если } (u, v) \notin E \end{cases}$$

Графы $G_1 = (V_1, E_1)$ и $G_2 = (V_2, E_2)$ называются изоморфными, $G_1 \cong G_2$, если существует биекция $\gamma : V_1 \to V_2$, сохраняющаяя ребра, т.е. для любых вершин $u_1, v_1 \in V_1$имеет место эквивалентность

$$\left(u_1^{\gamma}, v_1^{\gamma}\right) \in \mathrm{E}_2 \Leftrightarrow (\mathrm{u}_1, v_1) \in \mathrm{E}_1 \quad (3.1.1)$$

где u_1^γ и v_1^γ – образы вершин u_1 и v_1 относительно биекции γ. Последняя называется изоморфизмом графа G_1 на граф G_2. Множество таких изоморфизмов обозначим через $Iso(G_1, G_2)$. Таким образом,

$$G_1 \cong G_2 \Leftrightarrow \mathrm{Iso}(G_1, G_2) \neq \emptyset.$$

Легко видеть, что изоморфные графы имеют одинаковое число вершин и одинаковое число рёбер. Более того, поскольку, очевидно, изоморфизмы сохраняют степени вершин, то граф, изоморфный регулярному, сам регулярен и имеет ту же степень.

В таком случае может вознинуть вопрос, почему задача поиска DLP-системой защищаемого факта в передаваемом сообщении сводится к задаче поиска изоморфизма двух графов? Как правило, фильтруемые DLP-системой сообщения меньше, и даже существенно меньше, чем документы, которые содержат защищаемые факты. Поэтому точнее было бы говорить о задаче поиска подграфа в графе. А задача поиска подграфа в графе является NP-полной [59].

Проблема изоморфизма графов состоит в нахождении наиболее эффективного алгоритма, распознающего, являются ли два заданных графа изоморфными. Более точно, мы интересуемся вычислительной сложностью следующей задачи:

$\mathrm{Iso}(G_1, G_2)$: для графов G_1 и G_2 определить верно ли, что $G_1 \cong G_2$.

Не умаляя общности мы всегда будем предполагать, что оба графа имеют одинаковое число вершин n. Поэтому легко проверить, принадлежит ли данная биекция $\gamma : V_1 \to V_2$ множеству $\mathrm{Iso}(G_1, G_2)$: проверка условия (3.1.1) эквивалентна проверке n^2 равенств

$$(A_1)_{uv} = (A_2)_{u^\gamma v^\gamma}, \qquad u, v \in V_1$$

где $A_1 = A_1(G_1)$ и $A_2 = A_2(G_2)$. Поэтому проблема изоморфизма принадлежит классу NP. Однако, до настоящего времени неизвестно, принадлежит ли она классу P проблем, для которых существует алгоритм полиномиальной сложности; неизвестно также, является ли она NP-полной проблемой [38].

Также из [38] известно, что изоморфизм n вершинных графов распознаваем за время

$$\theta = \exp\left(O\left(\sqrt{n \log n}\right)\right) \quad (3.1.2)$$

Рассмотрим практические реализации задачи изоморфизма подграфа. Наиболее широко используемым алгоритмом решения этой задачи является алгоритм Ульмана [39]. В работе [61] показано, что этот алгоритм в сравнении с другими дает лучшие результаты с точки зрения временной сложности для задачи поиска изоморфизма пар графов.

В работе [39] представлен еще один быстрый алгоритм – VF2. Сравнение временной сложности этих алгоритмов приведено в табл. 3.1.1.

	Алгоритм VF2	**Алгоритм Ульмана**
Лучший случай	$O(N^2)$	$O(N^3)$
Худший случай	$O(N!\,N)$	$O(N!\,N^2)$

Таблица 3.1.1. Временная сложность алгоритмов поиска подграфов в графах.

Из приведенной выше табл. 3.1.1 видно, что временная сложность алгоритма поиска подграфов в графе достигает $O(N!\,N)$.

Как уже упоминалось выше, для построения семантической модели текстов естественного языка сейчас, как правило, используются графы. При необходимости реализовать поиск в графе (графе семантического описания защищаемого документа) подграфа (подграфа передаваемого сообщения) прийдется использовать один из обозначенный выше алгоритмов. В таком случае в связи с большой временной сложностью этих алгоритмов DLP-система столкнется с проблемой производительности уже на небольших документах.

Чтобы показать это рассмотрим документ на русском языке, состоящий из двух сраниц шрифта «Times New Roman», размер шрифта 12. По грубой оценке в таком документе содержится около 100 предложений, в среднем из семи слов. После синтаксического анализа можно ожидать семантический граф размерностью порядка 300 узлов. В случае большого количества циклов в графе необходимо рассматривать «худший» случай для оценки временной сложности

алгоритма поиска подграфа в графе. В этом случае временная сложность будет оцениваться числом 10^{617}. При использовании современных средств вычислительной техники операции такой сложности будут выполняться недопустимо долго, что делает невозможным использование алгоритмов поиска пографов в графах для идентификации защищаемых данных в передаваемых сообщениях.

Предлагаемый в данной работе метод идентификации защищаемых данных в передаваемых сообщениях основан на анализе связей между словами, и не предполагает построения дереьев. С точки зрения вычислительной сложности его можно описать следующим образом.

1. Формирование последовательности связей L_t мощностью $3 \cdot N_t$, где N_t – число вершин графа синтаксического описания передаваемого соощения.
2. Формирование последовательности L'_t на основе полученного на предыдущем шаге множества с учетом синонимов и дополнения нулями.
3. Вычисление функции корреляции (2.5.5.1) для $n = 0$.

$$F_{match}(L, L_t) = \sum_{i=0}^{N} L'_i \cdot L'_{ti},$$

где $N = |L'_t|$, а L'_i – последовате связей защищаемого документа.

Оценим временную вычислительную сложность предлагаемого метода.

Вычислительная сложность первого шага θ_1 оценивается слудующим образом.

$$\theta_1 = O(N_t) \text{ (3.1.3)}$$

Вычислительная сложность θ_2 второго шага не зависит от числа вершин N_t, и является константой.

$$\theta_2 = O(1) \text{ (3.1.4)}$$

Вычислительная сложность θ_3 третьего шага, как видно из (2.5.5.2), линейно зависит от N.

$$\theta_3 = O(N_t) \text{ (3.1.5)}$$

Из описания алгоритма формирования последовательностей L'_t и L' видно, что их мощности существенно больше, чем число связей, на основании которых они изначально сформированы. Если число вершин N_t , то число связей можно грубо оценить в $3 \cdot N_t$, а число связей с учетом синонимов в $3 \cdot 20 \cdot N_t$.Однако эти мультипликаторы не зависят от числа вершин N_t и поэтому опущены. Также важно отметить, что при оценке вычислительной сложности третьего шага θ_3 опущены заведомо безсмысленные умножения на нули.

Таким образом, вычислительную сложность предлагаемого метода можно оценить следующим образом.

$$\theta = O(N_t) \quad (3.1.6)$$

Важной особенностью является то, что сложность предложенного метода зависит от числа вершин N_t меньшего из семантических графов, а не большего как в случае с аглоритмами поиска подграфов в графах. Как правило, это семантический граф передаваемого сообщения.

	Предложенный метод	**Алгоритм VF2**
Лучший случай	$O(N_t)$	$O(N^2)$
Худший случай	$O(N_t)$	$O(N!N)$

Таблица 3.1.2. Временная сложность алгоритмов при решении задачи идентификации защищаемых данных в передаваемых сообщениях.

Из таблицы 3.1.2 видно, что представленный метод идентификации защищаемых данных в передаваемых сообщениях с точки зрения производительности существенно эффективнее, чем сравнение графов семантических деревьев. Это позволяет использовать его в DLP-системах с большим числом защищаемых документов без существенной задержки передаваемых сообщений на время анализа.

3.2 Оценка применимости предложенных решений

Оценка применимости метода снижения числа ошибок первого и второго рода в морфологических анализаторах DLP-систем

Хорошо известно, что на этапе автоматического морфологического анализа порождается морфологическая омонимия: в силу случайных совпадений парадигм разных частей речи некоторые словоформы интерпретируются неоднозначно [68].

Предлагаемый метод снижения числа ошибок первого и второго рода в морфологических анализаторах DLP-систем используется на этапе морфологического анализа лингвистического анализатора DLP-системы. В формальной модели DLP-системы этап морфологического анализа F_{22} (2.2.5) является одним из первых. В связи с этим для дальнейшего анализа важно получить высокие показатели точности на этом этапе.

Предлагаемый метод может использоваться как самостоятельно, так и в сочетании с другими методами, например с методом скрытой марковской модели (HMM - Hidden Markov Model), или с синтаксическим анализатором именных групп [69].

В работе [67] показано, что применение анализаторов, построенных на разных принципах, позволяет в автоматическом режиме снизить уровень неоднозначности до нескольких процентов, а в интерактивном режиме служит мощным средством контроля и отладки.

Применительно к морфологическим анализаторам DLP-систем больший интерес представляет автоматический режим. В связи с этим, предложенный метод эффективнее применять в сочетании с другими, основанными на других принципах. Это позволит сократить число гипотез морфологических описаний слов в предложениях анализируемого сообщения, что приведет повышению показателей качества полноты и точности выявления DLP-системой угрозы утечки конфиденциальной информации.

При этом предложенный метод обладает важным преимуществом – для его работы не требуются предварительно размеченные тексты или иные специально

подготовленные данные, за исключением словарей морфологических описаний слов.

Оценка применимости метода предотвращения передачи конфиденциальных ЕЯ собщений, содержащих отсутствующие в словаре парадигмы слов

В процессе создания базы знаний наиболее трудоемкой является процедура пополнения ее новой информацией, извлекаемой из специальных естественно-языковых текстов: деловых, научных, технических, медицинских, юридических и т. п. Очевидно, что при этом, в первую очередь, соответствующими терминами должны быть пополнены морфологический и семантический словари, на основе которых и выполняется анализ текстов.

Возможны два подхода к организации словарей, обеспечивающих их пополнение. Первый подход заключается в использовании дополнительных словарей пользователя, что позволяет хранить информацию в основных словарях в неизменяемом упакованном формате. Второй подход требует хранения всей информации в виде, допускающем коррекцию, пополнение и удаление словарных статей [70]. Особенности представления информации при таком подходе достаточно подробно рассмотрены в [71].

При введении DLP-системы в эсплуатацию происходит анализ всех защищаемых документов. В результате этого анализа выявляются слова, отсутствующие в словаре морфологических описаний слов. Для повышений показателей качества выявления угроз утечки конфиденциальной информации, содержащейся в проанализированных документах, необходимо пополнить словарь морфологических описаний слов найденными терминами. Предлогаемый метод позволяет существенно ускорить и упростить этот процесс.

Также предложенный метод может использоваться при периодическом пополнении множества защищаемых документов. В результате анализа каждого добавляемого документа возможно появление новых терминов, отстутствующих в

словаре морфологических описаний слов. В этом случае, аналогично предыдущему, также применим предложенный метод.

Наиболее важным в рамках DLP-систем применением описанного метода является анализ сообщений ЕЯ, содержащих отсутствующие в словаре парадигмы слов. Описанный метод позволяет автоматически получить морфологическое описание несловарного термина в анализируемом сообщении и пополнить морфологический словарь всеми его словоформами. Благодаря этому DLP-система может более корректно анализировать характерные для современных ИС ЕЯ-сообщения. Также появляется возможность уйти от последовательного внесения в морфологический словарь всех возможных словоформ с их морфологическими характеристиками, что является необходимой, но нетипичной задачей для служб ИБ и ИТ.

При этом остается достаточно много возможностей для доработки метода. В частности, видится полезным учет знаков препинания и отдельных словосочетаний [34].

Оценка применимости метода идентификации защищаемых данных в передаваемых сообщениях на основе анализа связей в объектной модели естественного языка

Описанный в работе метод идентификации защищаемых данных в передаваемых сообщениях на основе анализа связей в объектной модели естественного языка применяется на последнем этапе анализа, после семантического (Рисунок 3.2.1).

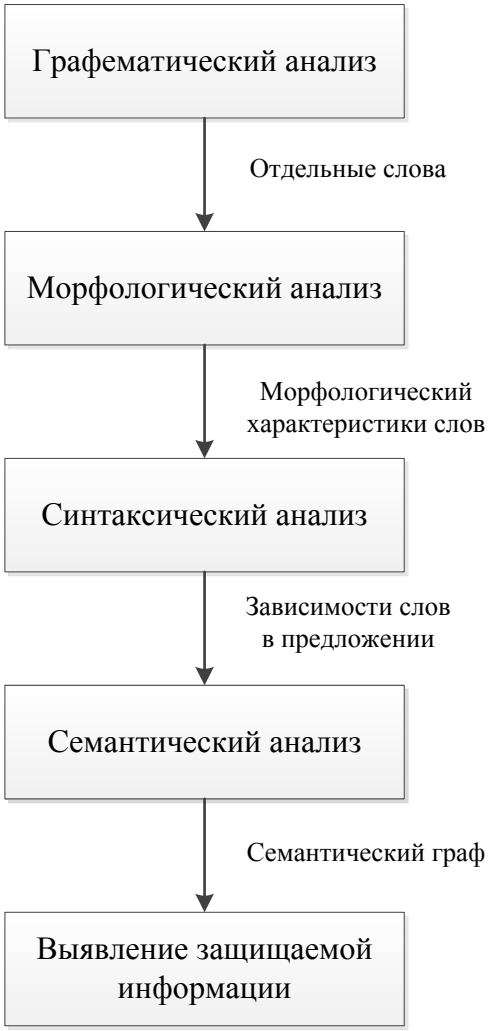

Рис. 3.2.1. Место метода идентификации защищаемых данных в передаваемых сообщениях на основе анализа связей в объектной модели естественного языка в функции морфологического анализа F_2

В терминах формальной модели DLP-системы это этап F_{25} (2.2.5) функции морфологического анализа F_2.

В главе 2 уже упоминалось, что на уровне синтаксического анализа с большой вероятностью будет иметься множество различных гипотез описания передаваемого сообщеия и защищаемых данных.

При анализе естественноязыковых сообщений с целью выявления угроз ИБ нет ничего плохого в том, что будут обнаружены неестественные или неочевидные трактовки передаваемого текста: если передаваемое сообщение можно тракотовать как защищаемый текст, то не важно, что именно имел ввиду

отправитель. Система защиты от утечек должна безусловно блокировать передачу такого сообщения.

Таким образом, неверная (точнее, непредполагаемая), но формально допустимая интерпретация передаваемого сообщения не может вызвать ошибку при выявлении утечек информации – «ложное» срабатывание в данном случае не является ложным.

Описанный в работе метод идентификации защищаемых данных в передаваемых сообщениях позволяет позволяет работать с множетсвом гипотез, поступающис после этапа семантического анализа. Полученные гипотезы должны быть последовательно обработаны, и при определении схожести хотя бы одной из них с защищаемыми данными передача сообщения должна быть заблокирована.

Однако следует учитывать особенности ЕЯ, носители которого обладают такими возможностями, до моделирования которых науке еще нужно пройти очень большой путь. Это, прежде всего, видение мира. За текстами ЕЯ человек видит картины внешнего мира, которые несут гораздо больше информации, чем самтекст. Человек способен по отдельным компонентам, присутствующим в тексте, восстанавливать эти картины, дополнять их, использовать причинно-следственные зависимости для прослеживания последующих изменений, динамики. Такая возможность выходит далеко за рамки моделей, основанных на логическом выводе. Отсюда следует особенность текстов ЕЯ. Как правило, в них умалчивается то, что известно адресатам, для которых предназначен текст, и что легко восстанавливается по тексту. Другими словами, большое количество нужной пользователю информации дается в текстах ЕЯ в скрытом виде. Такая информация называется имплицитной [62].

Надо отметить, что описанный метод идентификации защищаемых данных в передаваемых сообщениях не учитывает имплицитную информацию, которая может содержаться в тексте. Это является одним из перспективных направлений для дальнейшего развития предлагаемого метода.

Моделирование понимания в рамках интерактивного подхода (учитывающего действия участников общения) не ограничивается

распознаванием формы, дешифровкой (определением семантики частных значений лексенмы, так называемых лексико-семантических вариантов), «сложением», т.е. синтезом возможного смысла сообщения. Действия слушающего, как теперь очевидно, включают в себя и возможные выводы (импликатуры) из сказанного, и перебор возможных вариантов понимания с учетом «угадывания» намерений говорящего («если бы говорящий имел в виду X, он бы скорее сказал X, а не Y, как тут») и ряд других действий. Такая более громоздкая модель может применяться не только для моделирования поведения участников общения в «нормальных», тривиальных условиях, когда говорящий стремится к наиболее полному и однозначному пониманию («что имел в виду, то и сказал», т.е. понимать надо буквально).

Наиболее простой случай отклонения от буквального понимания – это, видимо, намек. Это высказывание, которое содержит в себе некоторые инструкции по формированию импликатур, желательных для говорящего. Отсутствие иллокутивной ценности высказывания («Зачем об этом говорить?») должно заставить адресата сделать возможные выводы.

Как мы уже отмечали, импликатуры обязательно сопровождают понимание сообщения [63]. В случае намека говорящий так выстраивает свои высказывания, что они не имеют ценности сами по себе и этим стимулируется действие слушающего. Заметим, что для понимания намека необходим полный набор информации, сопровождающий каждый речевой акт: понимание языковых единиц, знание контекста и наличие общих сведений у участников общения, правила выводов, для намеков с перлокутивными целями – представления об иерархии, речевом этикете (какие-то просьбы неприлично высказывать в лоб) и т.п. Как отмечалось в [64], информативность намека является следствием постулатов Грайса, а именно, принципа релевантности [65], хотя нарушаются многие другие постулаты. В той или иной степени, это можно отнести и к другим случаям нетривиальной подачи информации.

Остановимся на наиболее парадоксальном случае использования языковых средств – на иронии. По Квинтиллиану, ирония – это высказывание, которое надо

понимать в противоположном смысле. Однако в русском языке слово ирония, ироничный используется гораздо шире. Анализ бытовых употреблений слова ирония показывает, что в русском языковом сознании к иронии относится и насмешка, не связанная с «обратным» пониманием [66].

Предложенный метод практически не учитывает возможность нетривиального и понимания передаваемого сообщения. Несущественный эффект в этом направлении может быть достигнут использованием расширенных множеств синонимов, однако это не даст качественного решения проблемы небуквального понимания передаваемого сообщения. Это направление также является перспективным. Однако, в отличие от описанной выше проблемы передачи имплицитной информации, решение будет лежать вне описанного метода выявления защищаемых данных, поскольку он изначально предполагает буквальное понимание анализируемых тектов.

Необходимо отметить, что несмотря на указанные проблемы передачи имплицитной информации и нетривиального понимания передаваемых сообщений предложенный метод идентификации защищаемых данных в передаваемых сообщениях позволяет морфологическим анализаторам DLP-систем эффективно анализировать поступающие после семантического анализа гипотезы передаваемых сообщений. Это позволяет выявлять угрозу конфиденциальности защищаемой информации даже в случае предачи защищаемых фактов в измененной формулировке.

3.3 Выводы

1. Реализован описанный метод снижения числа ошибок первого и второго рода в морфологических анализаторах DLP-систем. На основе этой реализации был проведен эксперимент, показавший эффективность предложенного метода. Таким образом, использование метода в DLP-системе позволяет повысить показатели полноты и точности обнаружения утечки конфиденциальной информации.

2. Реализован описанный метод предотвращения передачи конфиденциальных ЕЯ собщений, содержащих отсутствующие в словаре парадигмы слов. Эксперимент показал, что использование предложенного метода позволяет повысить точность определения естественноязыковых конструкций в DLP-системах. Таким образом, использование метода в DLP-системе позволяет повысить показателей полноты и точности обнаружения утечки конфиденциальной информации.

3. Предложенный метод выявления защищаемых данных на основе анализа связей в объектной модели естественного языка, в отличие от предыдущих двух методов, проверить экпериментально чрезвычайно затруднительно. Показано, что представленный метод идентификации защищаемых данных в передаваемых сообщениях с точки зрения производительности существенно эффективнее чем методы, основанные на сравнении семантических графов. Временная сложность предложенного метода линейно зависит от числа вершин семантического графа, в отличие от квадратичной (в лучшем случае) или факториальной (в худшем случае) сложности методов, основанных на сравнении семантических графов. Это позволяет использовать предложенный метод в DLP-системах с большим числом защищаемых документов без существенной задержки передаваемых сообщений на время анализа.

Важной особенностью предложенного метода является то, что его сложность зависит от числа вершин меньшего из семантических графов (как правило, это семантический граф передаваемого сообщения), а не большего, как в случае с аглоритмами поиска подграфов в графах.

4. Предложенный метод определения морфологических характеристик слов эффективнее применять в сочетании с другими методами, основанными на других принципах. Это позволит сократить число гипотез морфологических характеристик слов предложений анализируемого сообщения, что приведет повышению показателей качества полноты и точности выявления DLP-системой угрозы утечки конфиденциальной информации. При этом предложенный метод обладает важным преимуществом – для его работы не требуются предварительно

размеченные тексты или иные специально подготовленные данные, за исключением словарей морфологических описаний слов.

5. Описанный метод предотвращения передачи конфиденциальных ЕЯ собщений, содержащих отсутствующие в словаре парадигмы слов, как это следует из его названии, применим в первую очередь для анализа сообщений, некоторые слова которых отсуствуют в словаре морфологических описаний слов.

Также предложенный метод может использоваться при периодическом пополнении множества защищаемых документов. В результате анализа каждого добавляемого документа возможно появление новых терминов, отстутствующих в словаре морфологических описаний слов. Для повышений показателей качества выявления DLP-ситемой угроз утечки конфиденциальной информации, содержащейся в защищаемых документах, необходимо пополнить словарь морфологических описаний слов найденными терминами. Предлогаемый метод позволяет существенно ускорить и упростить этот процесс.

6. Описанный в работе метод идентификации защищаемых данных в передаваемых сообщениях на основе анализа связей в объектной модели естественного языка применяется на последнем этапе анализа, после семантического.

На уровне синтаксического анализа с большой вероятностью будет иметься множество различных гипотез описания передаваемого сообщеия и защищаемых данных. Описанный в работе метод идентификации защищаемых данных в передаваемых сообщениях позволяет позволяет работать с множетсвом гипотез, поступающис после этапа семантического анализа. Полученные гипотезы должны быть последовательно обработаны, и при определении схожести хотя бы одной из них с защищаемыми данными передача сообщения должна быть заблокирована.

Важно отметить, что при анализе естественноязыковых сообщений с целью выявления угроз ИБ нет ничего плохого в том, что будут обнаружены неестественные или неочевидные трактовки передаваемого текста: если передаваемое сообщение можно тракотовать как защищаемый текст, то не важно,

что именно имел ввиду отправитель. Система защиты от утечек должна безусловно блокировать передачу такого сообщения.

Таким образом, неверная (точнее, непредполагаемая), но формально допустимая интерпретация передаваемого сообщения не может вызвать ошибку при выявлении утечек информации – «ложное» срабатывание в данном случае не является ложным.

Заключение

В ходе выполнения работы были получены следующие научные и практические результаты:

1. Исследование современных подходов к защите от утечек показало, что в сложившейся обстановке для эффективной работы DLP-решений необходима доработка и разработка новых методов анализа передаваемых данных.
2. Исследование показало, что существующие в настоящий момент модели ЕЯ нуждаются в доработке и приспособлению к нуждам лингвистических анализаторов DLP-систем.
3. В связи с этим, для повышения уровня защищенности DLP-систем необходимо разработать методы повышения качества анализа лингвистических анализаторов DLP-систем.
4. На основе формализованного описания современной защищаемой информационной системы разработана модель угрозы утечки конфиденциальной информации.
5. Приведено общее описание DLP-системы, на основании которого разработана формальная модель работы DLP-системы.
6. Анализ полученной модели угрозы утечки конфиденциальной информации в сочетании с формальной моделью работы DLP-системы показал возможные направления развития существующих методов анализа естественноязыковых сообщений в DLP-системах.
7. В рамках выбранных направлений развития морфологических анализаторов DLP-систем был разработан метод снижения числа ошибок первого и второго рода в морфологических анализаторах DLP-систем.
8. В рамках выбранных направлений развития морфологических анализаторов DLP-систем был разработан метод предотвращения передачи конфиденциальных ЕЯ собщений, содержащих отсутствующие в словаре парадигмы слов.
9. В рамках выбранных направлений развития морфологических анализаторов DLP-систем был разработан метод идентификации защищаемых

данных в передаваемых сообщениях на основе анализа связей в объектной модели естественного языка.

10. Реализован описанный метод снижения числа ошибок первого и второго рода в морфологических анализаторах DLP-систем. На основе этой реализации был проведен эксперимент, показавший эффективность предложенного метода. Таким образом, использование метода в DLP-системе позволяет повысить показатели полноты и точности обнаружения утечки конфиденциальной информации.

11. Реализован описанный метод предотвращения передачи конфиденциальных ЕЯ собщений, содержащих отсутствующие в словаре парадигмы слов. Эксперимент показал, что использование предложенного метода позволяет повысить точность определения естественноязыковых конструкций в DLP-системах. Таким образом, использование метода в DLP-системе позволяет повысить показателей полноты и точности обнаружения утечки конфиденциальной информации.

12. Предложенный метод выявления защищаемых данных на основе анализа связей в объектной модели естественного языка, в отличие от предыдущих двух методов, проверить экпериментально чрезвычайно затруднительно. Показано, что представленный метод идентификации защищаемых данных в передаваемых сообщениях с точки зрения производительности существенно эффективнее чем методы, основанные на сравнении семантических графов. Временная сложность предложенного метода линейно зависит от числа вершин семантического графа, в отличие от квадратичной (в лучшем случае) или факториальной (в худшем случае) сложности методов, основанных на сравнении семантических графов. Это позволяет использовать предложенный метод в DLP-системах с большим числом защищаемых документов без существенной задержки передаваемых сообщений на время анализа.

Важной особенностью предложенного метода является то, что его сложность зависит от числа вершин меньшего из семантических графов (как

правило, это семантический граф передаваемого сообщения), а не большего, как в случае с аглоритмами поиска подграфов в графах.

13. Предложенный метод определения морфологических характеристик слов эффективнее применять в сочетании с другими методами, основанными на других принципах. Это позволит сократить число гипотез морфологических характеристик слов предложений анализируемого сообщения, что приведет повышению показателей качества полноты и точности выявления DLP-системой угрозы утечки конфиденциальной информации. При этом предложенный метод обладает важным преимуществом – для его работы не требуются предварительно размеченные тексты или иные специально подготовленные данные, за исключением словарей морфологических описаний слов.

14. Описанный метод предотвращения передачи конфиденциальных ЕЯ собщений, содержащих отсутствующие в словаре парадигмы слов, как это следует из его названии, применим в первую очередь для анализа сообщений, некоторые слова которых отсуствуют в словаре морфологических описаний слов.

Также предложенный метод может использоваться при периодическом пополнении множества защищаемых документов. В результате анализа каждого добавляемого документа возможно появление новых терминов, отстутствующих в словаре морфологических описаний слов. Для повышений показателей качества выявления DLP-ситемой угроз утечки конфиденциальной информации, содержащейся в защищаемых документах, необходимо пополнить словарь морфологических описаний слов найденными терминами. Предлогаемый метод позволяет существенно ускорить и упростить этот процесс.

15. Описанный в работе метод идентификации защищаемых данных в передаваемых сообщениях на основе анализа связей в объектной модели естественного языка применяется на последнем этапе анализа, после семантического.

На уровне синтаксического анализа с большой вероятностью будет иметься множество различных гипотез описания передаваемого сообщеия и защищаемых данных. Описанный в работе метод идентификации защищаемых данных в

передаваемых сообщениях позволяет позволяет работать с множетсвом гипотез, поступающис после этапа семантического анализа. Полученные гипотезы должны быть последовательно обработаны, и при определении схожести хотя бы одной из них с защищаемыми данными передача сообщения должна быть заблокирована.

Важно отметить, что при анализе естественноязыковых сообщений с целью выявления угроз ИБ нет ничего плохого в том, что будут обнаружены неестественные или неочевидные трактовки передаваемого текста: если передаваемое сообщение можно тракотовать как защищаемый текст, то не важно, что именно имел ввиду отправитель. Система защиты от утечек должна безусловно блокировать передачу такого сообщения.

Таким образом, неверная (точнее, непредполагаемая), но формально допустимая интерпретация передаваемого сообщения не вызывает ошибку при выявлении утечек информации – «ложное» срабатывание в данном случае не является ложным.

Применение разработанных методов позволяет решить поставленную задачу повышения показателей качества защиты DLP-систем.

Литература

1. Вахонин С. Л. DeviceLock Endpoint DLP Suite – год на рынке DLP–систем / С.Л. Вахонин // «Information Security/ Информационная безопасность». – 2012. – № 2. – С. 34.

2. Здор В. DLP: двойная защита // «Information Security/ Информационная безопасность» – 2012. – № 3. – С. 27.

3. Васильев В. DLP в середине 2012 года // PC Week/RE. – 2012. – №14 (799). – С. 18.

4. Лебедев И.С., Борисов Ю.Б. Анализ текстовых сообщений в системах мониторинга информационной безопасности / И.С. Лебедев, Ю.Б. Борисов // Информационно–управляющие системы. – 2011. – № 2. – С. 37–43.

5. Кук Д., Бейз Г. Компьютерная математика. – М.: Наука. Гл. ред. физ.–мат. лит., 1990. – С. 261–269.

6. Перекрестенко А.А. Разработка системы автоматического синтаксического анализа на основе мягко контекстно–зависимой унификационной грамматики / А.А. Перекрестенко // Компьютерная лингвистика и интеллектуальные технологии: по материалам конференции «Диалог–2004». – (Т. 1. – Вып. 11.). – С. 81–92.

7. Joshi, A. K., Schabes, Y.: Thee Adjoining Grammars. // Handbook of Formal Languages. – 1997.– P. 69–123.

8. Bar–Hillel Y., Shamir E. Language and Information // Selected Essays on Their Theory and Application.– Addison–Wesley.– Reading, Mass., 1964. – P. 87–98.

9. Рабин М.О., Скотт Д., Конечные автоматы и задачи их разрешения / М.О. Рабин, Д. Скотт // Кибернетический сборник. – ИЛ. – 1962 – вып. 4 – С. 58–91.

10. Хомский Н. Три модели описания языка / Н. Хомский // Кибернетический сборник. – ИЛ. – 1961. – вып. 2 – С. 237–266.

11. Маркус С. Теоретико–множественные модели языков. – М.: Наука. Гл. ред. физ.–мат. лит., 1970. – С. 13–16.

12. Волкова И.А. Введение в компьютерную лингвистику. Практические аспекты создания лингвистических процессоров: Учебное пособие для студентов факультета ВМиК МГУ / И.А. Волкова. – М.: Издательский отдел факультета ВМиК МГУ, 2006. – С. 5–7.

13. Лапшин С.В., Лебедев И.В. Метод повышения точности автоматического определения частей речи слов предложения в морфологических анализаторах DLP–систем / С.В. Лапшин, И.В. Лебедев // Научно–технический вестник информационных технологий, механики и оптики. – СПб.– 2013. – № 4 (86) . – С. 124–128.

14. Еськова Н.А., Бидер И.Г., Большаков И.А., Формальная модель русской морфологии // Предварительные публикации Проблемной группы по экспер. прикл. лингвистике ИРЯ АН СССР, М., 1978.– 97 с.

15. С . О . Шереметьева , С . Ниренбу р г , 1996 Эмпирическое модели рование в вычислительной морфологии // НТИ, 1996.

16. J. Goldsmith. Unsupervised Learning of the Morphology of a Natural Language // University of Chicago, 1998.

17. Ножов И.М. Реализация автоматической синтаксической сегментации русского предложения: дисс. канд. технич. наук: 05.25.05 / Ножов Игорь Михайлович.– М., 2003.–148 с.

18. Белоногов Г.Г. Итоги науки и техники. Серия «Информатика».– 1984.– №8.

19. Курочкин Ю. InfoWatch набирает вес / Ю. Курочкин // IT News. – 2012.– №06.– С 4.

20. Исследование утечек информации и конфиденциальных данных из компаний и госучреждений России в 2012 году [Электронный ресурс] // Аналитический Центр InfoWatch. – Режим доступа: http://www.infowatch.ru/sites/default/files/report/analytics/russ/InfoWatch_rus_2012.pdf., свободный.– Загл. с экрана.

21. Джордж Ф. Люгер Искусственный интеллект: стратегии и методы решения сложных проблем, 4–е издание. – М.: Издательский дом «Вильямс». – 2003. – С. 575.

22. Цейтин Г.С. О соотношении естественного языка и формальной модели [Электронный ресурс]. – Режим доступа: http://www.math.spbu.ru/user/tseytin/nevformu.html

23. Ярочкин В.И. Информационная безопасность / В.И. Ярочкин.– М.: Академический Проект.–5–е изд. — 2008. – С. 18–26.

24. Левцов В., Зенин Н., Информационная безопасность. Система защиты от утечек информации / В. Левцов, Н. Зенин // Финансовая газета, январь 2009.

25. Кудинов А.С., Воропаев А.А., Калинин А.Л. Высокоточный метод распознавания концов предложений / Кудинов А.С. [и др.] // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог». – 2011. – С. 368 – 378.

26. Каневский Е.А. Некоторые вопросы пополнения морфологического словаря терминами предметной области / Е.А. Каневский // Труды Международного семинара Диалог 2001 по компьютерной лингвистике и ее приложениям. – М.: РосНИИ искусственного интеллекта.– 2001. – Т. 2. – С. 156–160.

27. Большаков И.А., Большакова Е.И. Автоматический морфоклассификатор русских именных групп // Компьютерная лингвистика и интеллектуальные технологии: по материалам конференции Диалог – 2012. – Т. 1. – Вып. 11.– С. 81–92.

28. Зализняк А.А. Грамматический словарь русского языка / А.А. Зализняк. – М.: Русский язык,– Изд. 4–е, испр. и доп.–1987.

29. Тузов В.А. Компьютерная семантика русского языка / В.А. Тузов. – СПб: Изд–во СПбГУ, 2004. – 400 с.

30. Боярский К.К., Каневский Е.А. Проблемы пополнения семантического словаря / К.К. Боярский, Е.А. Коневский // Научно–технический вестник СПбГУ ИТМО. – 2011. – № 2 (72). – С. 132–137.

31. Лапшин С. В., Лебедев И. С. Метод полуавтоматического формирования словаря морфологических описаний слов. // Научно–технический вестник информационных технологий, механики и оптики. – 2012. – № 5 (81). – С. 106–111.

32. Национальный корпус русского языка [Электронный ресурс]. – Режим доступа: http://ruscorpora.ru/corpora–usage.html, свободный. – Загл. с экрана.

33. Открытый корпус русского языка [Электронный ресурс]. – Режим доступа: http://opencorpora.org/dict.php, свободный. – Загл. с экрана

34. Кобзарева Т.Ю., Афанасьев Р.Н. Универсальный модуль предсинтаксического анализа омонимии частей речи в РЯ на основе словаря диагностических ситуаций / Т.Ю. Кобзарева, Р.Н. Афанасьев // «Компьютерная лингвистика и интеллектуальные технологии». Труды международного семинара «Диалог–2002». – М.: Наука, 2002. – Т. 2. – С. 258–268.

35. Manning C. D., Raghavan P., Schutze H. An Introduction to Information Retrieval. // Cambridge University Press, Cambridge, England – 2009, – P. 3.

36. Баскаков С.И. Радиотехнические цепи и сигналы / С.И. Басков/ – М.: Высшая школа, 2000.– С. 87.

37. Романюк Ю.А. Дискретное преобразование Фурье в цифровом спектральном анализе: учебное пособие / Ю.А. Романюк.– М.: МФТИ, 2007. – 120с.

38. И.Н. Пономаренко Проблема изоморфизма графов: Алгоритмические аспекты (записки к лекциям) [Электронный ресурс] // Санкт–Петербургское отделение Математического института им. В.А. Стеклова октябрь–декабрь, 2010. – Режим доступа – http://logic.pdmi.ras.ru/csclub/sites/default/files/graph_isomorphism_ponomarenko_lecture_notes.pdf

39. Luigi P. Cordella, Pasquale Foggia, Carlo Sansone, and Mario Vento. A (Sub)Graph Isomorphism Algorithm for Matching Large Graphs // Pattern Analysis and Machine Intelligence – Volume:26 , Issue: 10 – 2004 – P. 1367 – 1372

40. Daniel Jurafsky, James H. Martin, Speech and Language Processing, Second Edition. – Prentice Hall; 2nd edition – 2008 – P. 545–570

41. Galitsky B. Machine learning of syntactic parse trees for search and classification of text – Режим доступа: http://www.sciencedirect.com/science/article/pii/S0952197612002552, платный, http://robingets.me/robinlabs/html/assets/whitepapers/mlSentParseTreeSearchClassifProblems_EAAI_R2.pdf, свободный.– Загл. с экрана.

42. Диковицкий В.В., Шишаев М.Г. Обработка текстов естественного языка в моделях поисковых систем // Труды Кольского научного центра РАН. Информационные технологии. – Апатиты, 2010. – Вып. 1. – С. 29–34.

43. Brin, S. The Anatomy of a Large–Scale Hypertextual Web Search Engine [Электронный ресурс] / Sergey Brin, Lawrence Page. – Режим доступа: http://infolab.stanford.edu/pub/papers/google.pdf, свободный.– Загл. с экрана.

44. Некрестьянов И.С. Латентно–семантический анализ: Введение в латентно–семантический анализ [Электронный ресурс] / И.С. Некрасов . – Режим доступа: http://meta.math.spbu.ru/~igor/papers/lsa–prg/node2.html, свободный.– Загл. с экрана.

45. Калиниченко А.В. Сущность проблемы анализа текста в полнотекстовых поисковых системах. Подходы и пути решения [Электронный ресурс]. – Режим доступа: http://www.jurnal.org/articles/2010/inf12.html, свободный.– Загл. с экрана.

46. Когаловский, М.Р. Перспективные технологии информационных систем / М.Р. Когаловский. –М.: Компания АйТи, 2003. – 288 с.

47. Солтон, Дж. Динамические библиотечно–информационные системы. – М.: Мир, 1979.

48. Лифшиц Ю. Модели информационного поиска [Электронный ресурс] / Ю. Лифшиц. – Режим доступа: http://yury.name/internet/03ianote.pdf, свободный.– Загл. с экрана.

49. Vakkari, P. eCognition and changes of search terms and tactics during task performance: Л longitudinal study. In RIAO’ 2000 Conference Proceedings, Content Based Multimedia Information, College de France, Paris, France, April 12–14, 2000; RIAO, Ed.; CID: Paris, 2000; Vol. 1, P. 894–907.

50. Гаврилова Т.А. Использование онтологии в системах управления знаниями [Электронный ресурс] / Т.А. Гаврилова. – Режим доступа: http://big.spb.ru/publications/bigspb/kni/use_ontolog y_m_suz.shtml, свободный.– Загл. с экрана.

51. Гаврилова Т.А. Базы знаний интеллектуальных систем / Т.А. Гаврилова, В.Ф. Хорошевский. – СПб.: Изд–во «Питер», 2001. – 382 с.

52. Gruber, T.R. A Translation Approach to Portable Ontology Specifications [Электронный ресурс]. – Режим доступа: http://tomgruber.org/writing/ontolingua–kai–1993.pdf, свободный.– Загл. с экрана.

53. Wielinga, B. Framework and Formalism for Expressing Ontologies / B. Wielinga etc.// ESPRIT Project 8145 KACTUS, Free University of Amsterdam Deliverable, DO1b.1, 1994.

54. Журавлев, С.В. УИС «РОССИЯ». Автоматическое тематическое индексирование полнотекстовых документов / С.В. Журавлев, Б.В. Добров // Материалы научно–практической конф. «Проблемы обработки больших массивов неструктурированных текстовых документов», 2001.

55. Осипов, Г.С. Семантический поиск в сети интернет средствами поисковой машины Exactus [Электронный ресурс] / Г.С. Осипов, И.А. Тихомиров, И.В. Смирнов. – Режим доступа: http://www.raai.org/cai–08/files/cai–08_exhibition_31.doc, свободный.– Загл. с экрана.

56. Золотова, Г.А. Коммуникативная грамматика русского языка / Г.А. Золотова, Н. К. Онипенко, М.Ю. Сидорова // Институт русского языка РАН им. В.В. Виноградова. – М., 2004. – 544 с.

57. Чаща синтаксического разбора для абзаца текста [Электронный ресурс]. – Режим доступа http://www.dialog–21.ru/digests/dialog2013/materials/pdf/GalitskyB.pdf, свободный.– Загл. с экрана.

58. Описание русских конструкций с внешним процессором в системе автоматической обработки естественного языка [Электронный ресурс]. – Режим доступа: http://www.dialog–21.ru/digests/dialog2013/materials/pdf/BogdanovAV.pdf, свободный.– Загл. с экрана.

59. Гэри М., Джонсон Д. Вычислительные машины и труднорешаемые задачи.– М.: Мир, 1982. С. 194.

60. Белоногов Г.Г. Об использовании принципа аналогии при автоматической обработке текстовой информации / Г.Г. Белоногов // Проблемы кибернетики. – М: Наука. – № 28.–1974 г.

61. B.T. Messmer, “Efficient Graph Matching Algorithms for Preprocessed Model Graphs,” PhD Thesis, Inst. of Computer Science and Applied Mathematics, Univ. of Bern, 1996.

62. Кузнецов И.П., СоминН.В.. Выявление имплицитной информации из текстов на естественном языке: проблемы и методы [Электронный ресурс] / И.П. Кузнецов, Н.В. Сомин. – Режим доступа: http://www.mathnet.ru/links/7ce1de8ede18b34653a72292c00a70ac/ia184.pdf, свободный.– Загл. с экрана.

63. Borisova E. Special Entities Used for Governing the Processes of Understanding/Under standing by communication, eds. E.Borisova, O.Souleimanova. Cambridge Scholars Publishers, 2013. – P. 95–103.

64. Кобозева И.М., Лауфер Н.И. Об одном способе косвенного информирования / И.М. Кобозева, Н.И. Лауфер // Известия АН СССР. Сер. лит. и яз. 1988. Т. 47, No 5. С. 462–470.

65. Wilson Deirdre. On Verbal Irony / Deirdre Wilson, Dan Sperber // Irony in Language and Thought: A Cognitive Science Reader. – New York Lawrence Erlbaum Associates, 2007. – P. 35–55

66. Борисова Е. Г., Пирогова Ю. К., Моделирование нетривиальных условий понимания сообщения (на примере иронии). Диалог 21. [Электронный ресурс].– Режим доступа: http://www.dialog–21.ru/digests/dialog2013/materials/pdf/BorisovaEG.pdf, свободный.– Загл. с экрана.

67. Боярский К.К., Каневский Е.А. Разработка инструментария для полуавтоматической морфологической разметки текста / К.К. Боярский, Е,А. Каневский // Труды международной конференции «Корпусная лингвистика – 2008». – СПб: СПбГУ, Факультет филологии и искусств, 2008. – С. 83–88.

68. Кобзарева Т.Ю. Конфликт грамматики и статистики (автоматический анализ русского предложения) / Т.Ю. Кобзарева // Тринадцатая национальная конференция по искусственному интеллекту с международным участием КИИ–2012 16–20 октября 2012 г., г. Белгород, Россия: Труды конференции. Т. 1. – Белгород: Изд–во БГТУ, 2012. – С. 285–292.

69. Сокирко А.В, Толдова С.Ю. Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка / А.В. Секирко, С.Ю. Толдова // Интернет–математика –2005. Автоматическая обработка веб–данных. М.: Яндекс, 2005. С. 80–94.

70. Каневский Е.А. Некоторые вопросы пополнения морфологического словаря терминами предметной области / Е.А. Каневский // Труды Международного семинара «Диалог–2001» по компьютерной лингвистике и ее приложениям. – М.: РосНИИ искусственного интеллекта, 2001. – Т. 2. – С. 156–160.

71. Поминов А.В. Некоторые вопросы организации пополняемых автоматических словарей // Труды Международного семинара «Диалог–97» по компьютерной лингвистике и ее приложениям. Москва: РосНИИ Искусственного Интеллекта, 1997.– С. 233–237.

72. Лебедев И.С. Методология обнаружения угроз нарушения информационной безопасности в открытых компьютерных сетях на основе функциональной модели естественного языка : дис. д–ра тенх. наук : 05.13.19 / Лебедев Илья Сергеевич. – СПб., 2012. – 246 с.

73. Аверченков В.И. Система формирования знаний в среде Интернет: монография [Электронный ресурс] / В.И. Аверченков [и др.].– М.: Флинта, 2011.– С. 107–109.

74. Глобальное исследование утечек конфиденциальной информации в 2013 году [Электронный ресурс] // InfoWatch. – 2014. – Режим доступа: http://www.infowatch.ru/sites/default/files/report/analytics/russ/InfoWatch_global_report_2013.pdf, свободный.– Загл. с экрана.