Как повысить точность детектирования конфиденциальной информации

Основные функции системы защиты данных

Современная система защиты данных представляет собой набор инструментов и технологий для предотвращения или контроля перемещения конфиденциальной информации за переделы информационной системы компании. Такая система перехватывает и анализирует потоки данных, пересекающих периметр защищаемой информационной системы. При выявлении в потоке данных конфиденциальной информации (инцидента) срабатывает защита – запускается процедура реагирования на инцидент, например, блокирование передачи, предупредительное сообщение отправителю или оповещение ответственного лица. Копия информации сохраняется в архиве.

Схематично работу системы защиты данных можно представить так:

Анализ

Основной функцией системы защиты данных (DLP) является обнаружение (детектирование) в информационных потоках компании данных, соответствующих определенным параметрам. Поэтому именно алгоритмы анализа информации являются ключевыми для успешной работы DLP-решения и надежной защиты корпоративных данных.

Перехваченная информация анализируется сначала по ее внешним признакам – формальным атрибутам, например, для электронного письма выясняется, кто его отправлял, куда, когда и др. Однако для четкой идентификации конфиденциальности информации такого анализа недостаточно. Поэтому вторым этапом является извлечение содержимого перехваченной информации и его контентный анализ.

Анализ формальных признаков и контентный анализ

Сочетание двух типов анализа: анализа формальных атрибутов перехваченной информации и ее содержания (контентного анализа) позволяет наиболее точно детектировать конфиденциальную информацию, предотвратить нарушение политик безопасности и повысить эффективность использования DLP-решения.

Существует несколько видов контентного анализа, такие как цифровые отпечатки, файловые метки, анализатор шаблонов, лингвистический анализ. Каждая из этих технологий имеет свои преимущества и ограничения.

В продуктах InfoWatch используется несколько технологий контентного анализа – анализ комплексных текстовых объектов (анализатор шаблонов), цифровые отпечатки и лингвистический анализ с поддержкой морфологии. Совместное использование нескольких технологий контентного анализа существенно повышает надежность выявления конфиденциальной информации и позволяет защитить данные в течение всего их жизненного цикла.

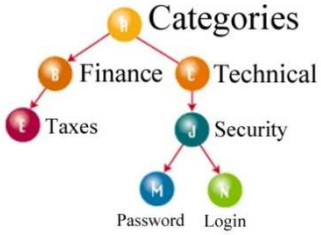
Лингвистический анализ и база контентной фильтрации

Технология лингвистического анализа позволяет автоматически определять тематику и степень конфиденциальности анализируемого фрагмента информации на основании встречающихся в нем терминов и их сочетаний.

Лингвистический анализ выполняется на основе заранее созданной базы контентной фильтрации (БКФ).

БКФ не только описывает категории информации, циркулирующей в компании, но и учитывает различные атрибуты её конфиденциальности, в т.ч. специфику деятельности компании, ее требования к безопасности. По результатам проведения лингвистического анализа тексту автоматически присваиваются те или иные категории, соответствующие его тематике и содержанию. В анализируемой информации могут встретиться термины (слова и словосочетания) из разных категорий, поэтому она может быть отнесена к одной или нескольким категориям БКФ.

База контентной фильтрации – это база данных, представляющая собой выделенный на основе вероятностных и математических методов иерархически организованный список (дерево) категорий с произвольным количеством вложенных уровней, и содержащая слова и выражения, наличие которых в документе позволяет определить тематику и степень конфиденциальности информации.



База контентной фильтрации и точность детектирования конфиденциальной информации

Надежность и точность идентификации конфиденциальных данных в корпоративных информационных потоках с помощью технологии лингвистического анализа зависят от базы контентной фильтрации, на основе которой осуществляется анализ.

Поэтому важно создать базу, которая обеспечит надежные результаты фильтрации информации по категориям. Основным методом лингвистического анализа с помощью БКФ является поиск в анализируемом фрагменте информации слов и словосочетаний, описывающих конфиденциальные данные и структурированных по категориям.

Создание БКФ

Для создания БКФ сначала нужно составить ее структуру – рубрикатор или дерево контентных категорий. Такое дерево представляет собой иерархический список с категориями и подкатегориями.

Затем каждую категорию нужно наполнить списком терминов, ключевых слов, словосочетаний и фраз, появление которых в анализируемом фрагменте информации указывает на его принадлежность к определенной контентной категории. После этого для каждого термина / словосочетания устанавливается вес, который этот термин будет иметь при отнесении информации к определенной категории. Решение о том, является ли текст релевантным контентной категории, принимается по результатам сравнения общей суммы веса терминов, найденных в тексте, с порогом релевантности этой категории.

Для обеспечения качественной категоризации базу контентной фильтрации необходимо поддерживать в актуальном состоянии – редактировать изменяющиеся со временем категории, добавлять и/или удалять термины и словосочетания, изменять их вес и др.

Характеристические и частотные термины

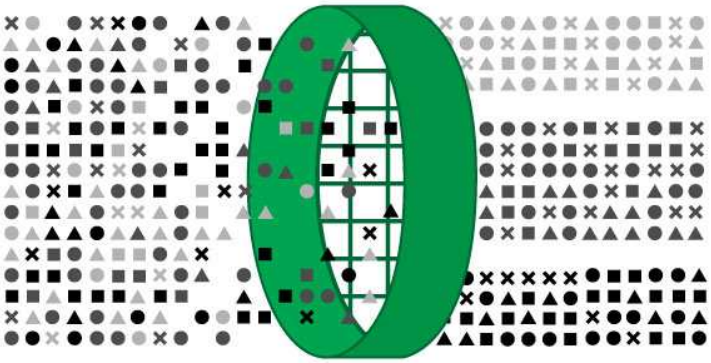
Термины, входящие в БКФ разделяются на частотные и характеристические.

Характеристический термин – это такой термин, который единожды встретившись в анализируемом фрагменте информации, 100% свидетельствует о принадлежности его к определенной категории.

Частотный термин – это такой термин, который, при наличии его в анализируемом фрагменте информации, с определенной долей вероятности свидетельствует о принадлежности этого фрагмента к определенной категории.

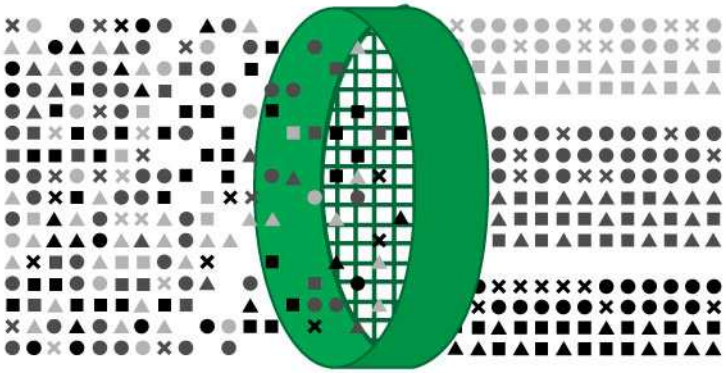
Базовая БКФ

В состав InfoWatch Traffic Monitor Enterprise входит стандартная база контентной фильтрации, содержащая наиболее общие категории и термины. Такая БКФ используется для демонстрации работы продукта и основных принципов лингвистического анализа в сети компании.



Отраслевая БКФ

На основе своего многолетнего сотрудничества с компаниями, работающими на различных вертикальных рынках, InfoWatch разработала несколько баз контентной фильтрации, оптимизированных под потребности конкретных сегментов рынка, например, финансовый, нефтегазовый, телекоммуникационный и др. секторы.



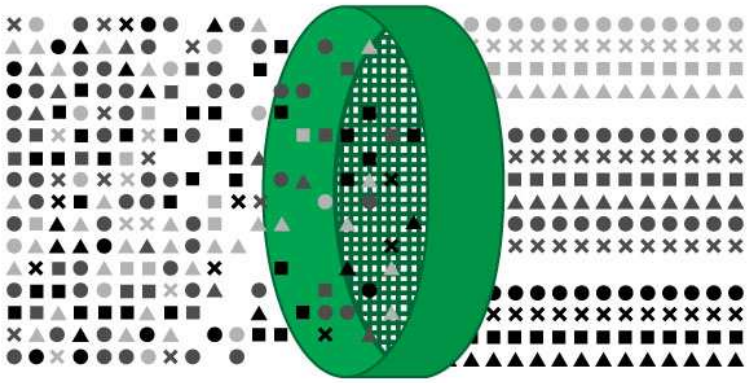
Оптимизация БКФ под какой-то сегмент рынка означает, что такая БКФ содержит наиболее распространенные категории, характерные для данной отрасли.

Запуск InfoWatch Traffic Monitor с предустановленной БКФ, оптимизированной под определенный вертикальный рынок, позволяет компании немедленно начать пользоваться продуктом и обеспечить примерно 60-70% точность детектирования конфиденциальной информации.

Почему важно адаптировать БКФ с учетом специфики деятельности конкретной компании

В базе контентной фильтрации, адаптированной под потребности определенной рыночной вертикали, примерно 80% категорий являются общими для всех компаний этого сектора.

Оставшиеся 20% составляют категории, характерные для конкретной компании.

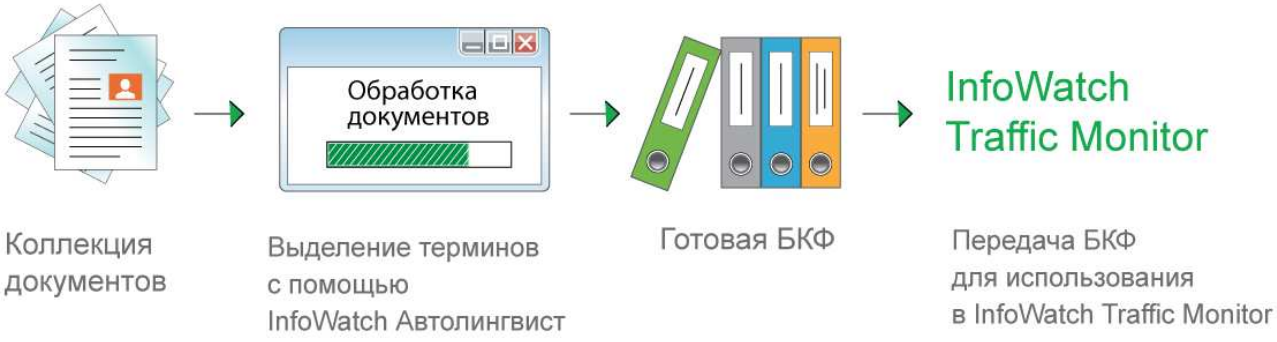


Дополнение такой отраслевой БКФ категориями, отражающими специфику деятельности данной компании, обеспечивает лучшую категоризацию и более точное детектирование конфиденциальной информации в информационных потоках компании – 70%+.

Дополнение БКФ специфическими категориями может происходить вручную, либо с использованием специального программного продукта – InfoWatch Автолингвист.

InfoWatch Traffic Monitor Автолингвист для создания собственной БКФ

InfoWatch Автолингвист – дополнительный программный продукт, предназначенный для совместного использования с InfoWatch Traffic Monitor, и позволяющий автоматизировать процесс создания собственной БКФ или доработки отраслевой БКФ. InfoWatch Автолингвист может быть использован для оценки качества полученной БКФ и поддержания ее в актуальном состоянии.
Для создания БКФ с помощью InfoWatch Автолингвист необходимо подготовить репрезентативную коллекцию документов компании и рассортировать ее по отдельным папкам в зависимости от тематики, например, финансовые документы, договора о неразглашении и др. После обработки коллекции документов с помощью InfoWatch Автолингвист эти папки составят структуру рубрикатора.
InfoWatch Автолингвист анализирует загруженную в него коллекцию документов и автоматически выделяет термины, на основании которых будет происходить отнесение анализируемой информации к той или иной категории.
Заключительным этапом при создании базы контентной фильтрации является добавление в нее характеристических терминов, которые не могут быть выделены автоматически, но одновременно однозначно свидетельствуют о конфиденциальности документа, такими как, например, название секретного проекта.



Использование отраслевой базы контентной фильтрации, дополненной специфическими для конкретной компании терминами и категориями, позволяет не только обеспечить точную категоризацию корпоративной информации, но и, в результате, существенно сократить затраты: дополнение отраслевой БКФ специфическими категориями с помощью InfoWatch Автолингвист занимает в 3 раза меньше времени, чем если бы это было сделано вручную.

Поддержание БКФ в актуальном состоянии тоже происходит в автоматическом режиме: не нужно выделять отдельного специалиста для анализа новых документов, выделения из них терминов и доработку структуры базы – это сделает InfoWatch Автолингвист.

БКФ под ключ

Компания InfoWatch работает на рынке защиты корпоративных данных с 2003 года. За это время специалисты InfoWatch внедрили продукты компании в сетях более 100 клиентов, помогая им создавать собственные базы контентной фильтрации и дорабатывая их, чтобы обеспечить максимально точную идентификацию конфиденциальных данных.

Вы можете воспользоваться уникальным опытом лингвистов InfoWatch при разработке собственной БКФ. В этом случае лингвисты InfoWatch возьмут на себя основную работу по выделению категорий и наполнению их терминами, предоставив вам готовую БКФ под ключ. Они также проведут курс обучения для ваших сотрудников по использованию базы и поддержанию ее в актуальном состоянии.

Виды БКФ и точность детектирования

Лингвистический консалтинг

Чтобы помочь клиентам максимально полно использовать весь функционал наших продуктов, добиться точной категоризации и высочайшего качества детектирования конфиденциальных данных, InfoWatch предлагает лингвистический консалтинг, который включает в себя:

- Обучение использованию InfoWatch Автолингвист
- Рекомендации по подбору документов, для составления наиболее полной базы контентной фильтрации
- Рекомендации по рубрикации документов и выделению характеристических терминов

Контакты:
тел.: **+7 495 22 900 22**
Российская Федерация, 123458, Москва, проезд №607, дом 30, офис 507
www.infowatch.ru
sales@infowatch.ru