

УДК 004.056

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ ЗАЩИТЫ САЙТОВ ОТ ИЗВЛЕЧЕНИЯ И ОБРАБОТКИ ДАННЫХ

Бабарицкий П.А. (Университет ИТМО)

Научный руководитель - к.п.н., доцент Государев И.Б.
(Университет ИТМО)

Настоящая статья охватывает актуальную проблему противодействия методам и средствам несанкционированного извлечения и обработки данных веб-ресурсов. В рамках исследования освещаются теоретические положения и практические аспекты противодействия различным средствам несанкционированного доступа к данным. Практическая часть публикации представляет собой сравнительный анализ существующих методов защиты данных сайтов от извлечения и обработки данных. Научная новизна настоящего исследования определяется отсутствием положений комплексного сравнительного анализа ряда существующих методов выявления и устранения фактов несанкционированного извлечения и обработки данных веб-ресурсов. Автором статьи изложены результаты аналитического обзора и сравнительного анализа на основе собственной введенной системой сравнения методов с последующей оценкой их эффективности по ряду характеристических критериев.

Анализ методов извлечения и обработки данных Интернет-ресурсов является актуальным направлением исследования в области прикладной веб-безопасности и коммерческого использования веб-ресурсов. Об этом свидетельствует множество современных научных публикаций, посвященных проработке положений указанной проблемной области. В таких работах раскрывается специфика и назначение такого явления как web-scraping, которое предполагает совокупность методик автоматического извлечения необходимых данных веб-ресурса с целью минимизации человеческого фактора в рамках этой задачи. Такие исследования позволяют создавать системы размещения Интернет-рекламы на основе единой системы фильтрации с использованием технологии web-scraping. Также в рамках исследований предлагается новые концепции спам-ботов, а также приводятся результаты экспериментальных исследований, определяющих модель поведения спам-ботов. Полученные результаты могут быть заложены в основу политики идентификации спам-ботов и фильтрации нежелательного веб-контента. Несмотря на достаточно большое количество научных трудов в области исследования методов обнаружения и борьбы с нежелательными ботами, текущее состояние изученности вопроса может быть улучшено за счет комплексного сравнительного анализа методов защиты веб-ресурсов от извлечения и обработки данных. Так, проблема настоящего исследования воплощается в потребности сравнительного анализа существующих методов борьбы с несанкционированным извлечением данных веб-ресурсов и построения системы оценки эффективности анализируемых методов по ряду характеристических критериев. Таким образом, на основании изложенной научной проблемы были определены объект и предмет исследования. Объект исследования: методы защиты сайтов от извлечения и обработки данных. Предмет исследования: методологии и технологии сравнительного анализа методов защиты сайтов от извлечения и обработки данных.

В рамках аналитического обзора рассматриваются существующие методы и средства по противодействию несанкционированному извлечению и обработке данных с веб-ресурса. По результатам аналитического обзора был выявлен ряд моделей, средств и методов противостояния воздействию нежелательных веб-роботов, который был разделен на две глобальные группы: базовые и высокоэффективные. Каждую из приведенных методологий предлагается подвергнуть краткому рассмотрению для раскрытия теоретических аспектов противодействия веб-роботам, нацеленным на несанкционированную обработку данных веб-ресурсов. Также данный этап представляет собой подготовительную стадию для проведения

сравнительного анализа выявленных методологий. Для задачи проведения сравнительного анализа необходимо ввести систему сравнения, на основе положений которой будет выполняться анализ. Ныне не существует специально определенной системы, нацеленной на анализ методов борьбы с веб-роботами, чем и определяется научная потребность и научная новизна настоящего исследования. Конечной целью сравнительного анализа является определение наиболее оптимального метода противодействия веб-угрозам. Предлагается анализировать методы по ряду характеристических критериев с формированием условного рейтинга относительного каждой из анализируемых методологий. Рейтинг метода будет сформирован на основе набора оценок от 1 до 5 баллов (где 1 – полное несоответствие требованию характеристики, 5 – максимальное соответствие требованию характеристики) по каждому характеристическому критерию. В рамках системы анализа для всех критериев вводятся следующие характеристические признаки: актуальность, доступность реализации, доступность использования, область воздействия, бюджет реализации, уровень доверия, уровень эффективности. Для получения конечных оценок по критериям был исследован реальный опыт применения методологий защиты веб-ресурсов на основе источников со свободным доступом. Кроме этого, были учтены результаты анализа достаточно известного платежного веб-ресурса – электронного кошелька paypal.com. Поскольку к Интернет-сервисам, работающим с денежными средствами, предъявляются наиболее жесткие требования безопасности, на сайте paypal.com реализованы сразу два метода защиты от несанкционированного извлечения данных: пользователя вынуждают пройти процедуру аутентификации и ввести набор символов. Полученные результаты сравнительного анализа показывают, что наиболее оптимальным для использования методов защиты веб-ресурсов является CAPTCHA. Это объясняет такую широкую распространенность настоящего метода в области защиты данных сайтов и её усовершенствования. Наиболее эффективной моделью защиты с точки зрения всех аспектов следует признать комбинирование методов. Например для более надежной защиты данных веб-ресурс может потребовать пользовательскую аутентификацию, проверку с помощью CAPTCHA в процессе ввода данных и быть защищено пользовательским соглашением об использовании веб-ресурса для обеспечения юридической защищенности.

В рамках проведенного исследования была в полной мере достигнута цель и успешно решены поставленные задачи в соответствии с объектом и предметом исследования. Был выполнен аналитический обзор существующих методов защиты веб-ресурсов от несанкционированного извлечения и обработки данных. По результатам аналитического обзора были изложены краткие теоретические положения о специфике и назначении метода. В рамках практической части исследования был выполнен сравнительный анализ методов защиты данных сайтов, рассмотренных в рамках аналитического обзора. По результатам проведенного анализа было установлено, что наиболее оптимальным методом защиты веб-ресурсов является CAPTCHA – hiding data in the media. Метод с самым низким показателем – Using robots.txt. Полученные результаты анализа не могут быть однозначны и полностью объективны, тем не менее, в рамках проведенного научного исследования сделана попытка оценки существующих методов противодействия веб-угрозам на основе введенной автором системы сравнительного анализа методов. Полученные результаты имеют практическую ценность и могут быть основой для дальнейших научных исследований методов противодействия веб-угрозам.