Анализ механизмов обнаружения запрещенного содержимого в сети Интернет

А.А. Фролов, Д.С. Сильнов, А.М. Садретдинов

***Аннотация*— В данной статье рассматривается проблема распространения запрещенного содержимого в сети Интернет, а именно исследование механизмов выявления этого содержимого на веб-страницах. Проводится анализ существующих решений в этой области, в том числе законодательные меры, обеспечивающие процесс удаления запрещенного содержимого с ресурса или ограничения доступа к нему, информация из отчетов организаций, ограничивающих доступ к запрещенным ресурсам или отдельным страницам этих ресурсов, которая свидетельствует о росте общего количества запрещенного контента в сети. Приводятся результаты анализа существующих методов обнаружения запрещенного содержимого на основании которых принимаются решения о необходимости разработки новых подходов к выявлению этого содержимого. Большое внимание уделяется проблеме защиты детей в сети Интернет, а также проблеме распространения запрещенного контента посредством скрытых сетей и другие угрозы пользователям сети, возникающие из-за этих сетей. Описывается эксперимент, который проводился для анализа содержимого веб-ресурсов с целью повышения качества механизмов выявления запрещенного содержимого в сети Интернет и сети Darknet. Во время проведения эксперимента для извлечения и обработки веб-страниц был разработан программный модуль, который производил разметку веб-страницы специальными маркерами для разделения содержимого веб-документа на группы. Полученные по итогам эксперимента данные содержат результаты тестирования разработанного модуля и примеры его работы. В заключительной части статьи приводится набор методов, которыми автор предполагает воспользоваться в дальнейших исследованиях, в которых планируется разработка программного комплекса принятия решений о наличии на веб-ресурсах сети Интернет запрещенного содержимого.**

***Ключевые слова*— Интернет; фильтрация; запрещенный контент.**

I. Введение

Данная статья является продолжением работ [1-2] по пресечению распространения запрещенного содержимого в сети Интернет. Запрещенным контентом считается любое содержательное наполнение информационного ресурса или веб-сайта, которое было запрещено государством для распространения или просмотра [3-4]. В предыдущих работах все экспериментальные исследования проводились в основном на ресурсах скрытой сети Darknet с основной целью – удалить запрещенное содержимое, которое распространяется на этих ресурсах. Для обнаружения ресурсов с запрещенным контентом в сети Darknet не обязательно разрабатывать сложные программы для сканирования и обработки информации с веб-страниц, так как большая часть всех ресурсов данной сети содержит и распространяет запрещенный контент [5-6]. Данное исследование призвано доработать недостатки методов, выделенных в предыдущих исследованиях или выявить новые.

В связи с тем, что во время исследования имел место анализ запрещенного контента, некоторые моменты исследования будут опущены, чтобы избежать нарушения законодательства.

II. Анализ проблемы

В Великобритании действует Internet Watch Foundation (IWF) – благотворительная организация, находящая в Графстве Кембридж, задача которой свести к минимуму наличие криминального содержимого в сети Интернет, а именно контента, содержащего насилие над детьми, преступно непристойный контент для взрослых, материалов, разжигающих расовую ненависть, преступную деятельность[1]. Эта и другие подобные организации из других стран преимущественно занимаются проблемами обеспечения детской безопасности в сети, так как число надругательств над детьми в киберпространстве увеличивается с каждым годом [7].

В 2015 году информация по 68092 сообщениям была идентифицирована IWF как содержащая детскую порнографию и была заблокирована. В 2016 году 59548, а в 2017 году цифра составила 80318. Этот показатель включает URL-адреса и группы новостей. Преступники все чаще используют методы маскировки, чтобы скрыть изображения и видео сексуального насилия над детьми в интернете и оставлять подсказки педофилам, чтобы они могли найти незаконный контент, скрытый за разрешенным. В 2016 году IWF нашла 1572 веб-сайта,

Статья получена 30 декабря 2018. Рекомендована организационным комитетом III Международной научной конференции «Конвергентные когнитивно-информационные технологии».

Фролов Алексей Алексеевич. Аспирант. Аспирант кафедры №12, Национальный исследовательский ядерный университет «МИФИ» (email:aleksey2093@outlook.com)

Сильнов Дмитрий Сергеевич. Кандидат технических наук. Доцент кафедры №12, Национальный исследовательский ядерный университет «МИФИ» (email: ds@silnov.pro)

Садретдинов Артем Маратович. Студент кафедры №12, Национальный исследовательский ядерный университет «МИФИ». (email:artemoner@mail.ru)

использующих метод, чтобы скрыть изображения сексуального насилия над детьми. Это на 112% больше, чем 743 замаскированных сайта, выявленных в 2015 году. В 2017 году число обнаруженных подобных веб-сайтов выросло еще на 86% до 2909. Это предполагает повышение уровня интеллекта среди правонарушителей, которые могут пойти на новые шаги, чтобы избежать обнаружения. Сайты хостинга изображений являются наиболее часто используемым среди преступников сервисом 78% от всего контента в 2015 году, 72% в 2016 и 69% в 2017. В Европе в настоящее время размещается большинство веб-страниц о сексуальном насилии над детьми, и с 2016 по 2017 год количество возросло с 60% до 65% от всего контента. Ведущими странами размещения URL-адресов с сексуальным насилием над детьми являются Нидерланды, США, Канада, Франция и Россия. В 2016 году 92% всех URL-адресов, связанных с сексуальным насилием в отношении детей, которые были определены во всем мире в 2016 году, были размещены в пяти странах: Нидерландах (37%), США (22%), Канаде (15%), Франции (11%) и России (7%)[2].

В статьях [8-10] большое внимание уделяется сложностям противодействия экстремизму в сети Интернет. Автор работы приводит различные примеры и практики Прокуратуры. Особое внимание уделяется обеспечению контент-фильтрации на территории образовательных учреждений, так как часть людей привлекаемых к уголовной ответственности за экстремизм несовершеннолетние. Кроме того, несовершеннолетних проще привлечь к различной экстремистской или другой деятельности, так как подростки могут не понимать, что их действия не законы [11].

Для фильтрации СМС сообщений чаще всего используются методы подсчета отправки одинаковых сообщений с одного или нескольких номеров. Так же операторы блокируют некоторых абонентов, если они отправляют слишком много сообщений в короткие промежутки времени, что, скорее всего, свидетельствует о том, что используется программа для рассылки спама посредством СМС сообщений. В работе [12] авторы предложили подход к обнаружению спама в тексте СМС сообщений с использованием машинного обучения. Кроме того, имеются проблемы с фильтрацией и ограничением доступа к нежелательной информации несовершеннолетними, которые решаются чаще всего локальными фильтрами на основе черно-белых списков и правил [13].

В социальных сетях существует большой риск распространения запрещенной информации или утечки конфиденциальных данных, например, фотографии, файлы из историй переписок пользователей, которые в последствии могут стать запрещенным контентом в руках злоумышленников [14]. В сети распространяется информация, побуждающая к каким-либо нежелательным действиям, например, суицидальный контент, инструкции к суициду и прочие [15]. Попадая в социальные сети подобный контент может очень быстро распространиться между группами пользователей.

Кроме распространения запрещенного контента, в сети можно обнаружить ресурсы, распространяющие запрещенные товары. Обычно такие ресурсы продают различные наркотические вещества, а в качестве способов оплаты используют криптовалюты. Именно из-за использования криптовалют спецслужбы практически не могут отслеживать деятельность таких площадок в скрытых сетях [16].

Для ограничения доступа к запрещенным сайтам на государственном уровне большинство стран используют так называемые черные списки. Черный список представляет из себя список ссылок, IP адресов, которые запрещены на территории страны. Различные сервисы Интернет-провайдеров могут проверить, если ли сайт, к которому пытается получить доступ пользователь в этом списке и, если результат положительный вместо страницы веб-ресурса пользователь увидит страницу с информационным сообщением, что данный ресурс заблокирован и причины блокировки. В России черный список это Единый реестр доменных имен, указателей страниц сайтов в сети «Интернет» и сетевых адресов, позволяющих идентифицировать сайты в сети «Интернет», содержащие информацию, распространение которой в Российской Федерации запрещено, далее ЧСРФ[3]. Обеспечением актуальности данных в ЧСРФ занимается Роскомнадзор[4]. Всего в ЧСРФ было занесено более 300 тыс. записей, в 2017 году количество записей менее 150 тыс. из них [17]:

- 40% - азартные игры;
- 37% - информация за распространение которой предусмотрена уголовная или административная ответственность;
- 11% - наркотические вещества;
- 8% - детская порнография;
- 4% - суицид.

Блокировке подвергаются не только сайты, нарушающие законодательство, но и сайты, распространяющие информацию о том, как попасть на запрещенные ресурсу [18]. Основные преимущества и недостатки ЧСРФ, которые можно выделить [19-23]:

- преимущества:
- если ресурс находится в черном списке, получить к нему доступ обычными способами не получиться;
- масштаб. Ограничение доступа к включенным в черный список ресурсов осуществляются всеми интернет-провайдерами, действующими на территории РФ;
- заявку о наличии на ресурсе запрещенного контента может оставить любой пользователь сети на официальном сайте ЧСРФ. В дальнейшем ресурс будет проверен сотрудниками Роскомнадзор на предмет наличия запрещенного контента и если такой контент присутствует на страницах сайта, то принимается

[1] https://www.iwf.org.uk/what-we-do/who-we-are (дата обращения 15.09.2018)

[2] https://www.iwf.org.uk/what-we-do/who-we-are/annual-reports (дата обращения 15.09.2018)

[3] http://eais.rkn.gov.ru (дата обращения 15.09.2018)

[4] http://rkn.gov.ru (дата обращения 15.09.2018)

решение о блокировки ресурса и через какое-то время ресурс появляется в черном списке.

• недостатки:

• владелец включенного в черный список и распространяющего запрещенный контент ресурса может обойти блокировку создав зеркало сайта, путём регистрации нового доменного имени. Несмотря на то, что по закону Роскомнадзор имеет право добавлять зеркала заблокированных сайтов в чёрный список без одобрения суда, зеркала многих ресурсов блокируются далеко не быстро. Многие торрент-трекеры и пиратские онлайн кинотеатры после блокировки Роскомнадзором основного адреса создают зеркало, которое работает ещё несколько месяцев прежде, чем новый адрес будет занесён в чёрный список. При создании зеркала владельцы обычно делают переадресацию с блокируемой страницы на новую, прежде чем основная будет заблокирована;

• не включенные в реестр ресурсы, содержащие запрещенный контент, не будут блокироваться. К таким ресурсам пользователи будут получать доступ до тех пор, пока ресурс не будут включен в реестр;

• исключить ресурс из черного списка при условии, что он попал туда случайно достаточно сложно, так как владельцу ресурса придётся доказывать, что его сайт попал в чёрный список по ошибке. Возможно, придется обращаться в суд;

• возможность обойти ограничение доступа к ресурсу при помощи прокси. Несмотря на закон о запрете обхода блокировок, часть пользователей сети получают доступ к заблокированным ресурсам, обычно через прокси-сервер, расположенный в другой стране, где Интернет-провайдер не блокирует доступ к этому сайту;

• после блокировки популярность ресурса может возрасти, так как пользователи могут по-прежнему получить к нему доступ, если обойти блокировку провайдера. Так после блокировки в 2016 году популярного торрент торрент-трекера, активность на сайте увеличилась почти в полтора раза, в то время как Роскомнадзор утверждал, что посещаемость, наоборот, упала. Как оказалось аудитория сайта после блокировки стала использовать прокси-сервера и таким образом на сайте могло быть 2 или более пользователей под одним IP-адресом[5].

III. Цель исследования

Для определения наиболее эффективных способов обнаружения запрещенного содержимого на веб-страницах было решено:

1. Проанализировать существующие методы и технологии, выявить их достоинства и недостатки;
2. Выбрать ряд ресурсов:
   a. Содержащих запрещенный контент;
   b. Не содержащих запрещенный контент;
   c. Ресурсы с информацией против запрещенного контента. Например, статья о вреде от наркотиков;
3. Сравнить распространяющие и не распространяющие запрещенное содержимое ресурсы. Определить, чем ресурсы первого типа отличаются от ресурсов второго типа;
4. Спроектировать алгоритм анализа содержимого страниц подобных ресурсов;
5. Реализовать программный модуль для решения задач: сканирования и обработки;
6. Сделать выводы и определить дальнейшие направления исследований.

IV. Проведение эксперимента

На первом этапе эксперимента был сформирован список содержащих запрещенный контент сайтов. Данные ресурсы были взяты из поисковых систем, ЧСРФ и скрытой сети Darknet. Включение в тестовый набор сайтов из сети Darknet позволило провести сравнение способов размещения и распространения контента на ресурсах и спроектировать более эффективные обработки содержимого веб-страниц. Сформированный список был проанализирован и в результате были выделены категории, по которым ресурсы были соответственно распределены: лента, форум, магазин, файлообменник. В дальнейшем список категорий будет пополняться новыми типами или подтипами сайтов для реализации более эффективных алгоритмов разделения содержимого веб-документов. Сравнение запрещённых и не запрещённых ресурсов показало, что способы размещения контента на страницах не различаются, а вот способы получения доступа и передачи, наоборот. Например, распространяющие запрещенные файлы ресурсы распространяют контент в архивах формата .rar с установленным паролем. Классический формат .zip не используется, скорее всего, по причине того, что он не позволяет полностью скрыть содержимое архива: после установки пароля пользователь может увидеть список файлов внутри архива не вводя пароль и догадаться о том, что находится внутри. По этой причине распространители предпочитают формат .rar, так как без пароля никто не сможет понять, что находится внутри архива [24]. В сети Darknet запрещенные файлы обычно распространяются без паролей, исключение составляют файлы, размещенные на ресурсах сети Интернет, так как высокая анонимность пользователей сети позволяет распространять запрещенный контент в открытом виде [25].

Для эффективного анализа ресурсов было решено создать программный комплекс анализа содержимого страниц и выдачу промаркированной страницы. Разделение содержимого страницы уже использовалось в некоторых решениях задач обнаружения запрещенного содержимого в сети Интернет. В работе [26] с помощью разметки на основе маркеров выделялось основное содержимое страницы с целью повышения качества выявления на сайте запрещенного контента по теме «Наркотики и наркомания», так же при оценке сайтов использовался метод жанровой классификации ресурсов [27]. В рамках текущего эксперимента было решено

[5] Администрация RuTracker признала падение посещаемости вдвое за год «вечной» блокировки. https://vc.ru/flood/21481-rutracker-admin-50percent [электронный ресурс] (дата посещения 17.08.2018).

выделять на страницах следующие типы информации:

• мета-теги – список всех мета-тегов, которые находятся в теге <head>;

• шапка сайта – заголовок страницы, пункты меню, краткое описание и другое, что может находится в верхней панели сайта;

• основное содержимое или контент страницы – все что находится внутри контейнера. Основное содержимое разделяется по вышеперечисленным категориям типа ресурса. В каждой категории содержимое дополнительно разделяется, например, в онлайн-магазинах есть страницы каталога товаров и страницы отдельных товаров. В первом случае страница будет содержать изображения и названия товаров, а во втором изображение товара, название, цену, описание, комментарии пользователей (набор полей зависит от ресурса);

• нижняя панель (toolbar) – копирайт сайта, счетчики поисковых систем для получения данных о посещаемости, ссылка на хостинг и т.д.;

• картинки – все изображения, которые есть на странице. Если изображения находятся внутри какого-либо элемента страницы, то между ними образуется иерархическая связь. Оценка изображения будет использоваться в алгоритме анализа шапки, основного содержимого, нижней панели;

• медиа содержимое – видео или аудио содержимое страницы, например, проигрыватель онлайн видео или музыки. В эту категорию будут определены JavaScript's, которые воспроизводят медиа файл;

• реклама – любые рекламные блоки, которые будут обнаружены на сайте;

• прочие – любое содержимое сайта, которое система не сможет определить ни в одну из вышеперечисленных категорий. Фрагменты веб-страниц, которые попали сюда, в дальнейшем будут проанализированы для доработки процесса сканирования содержимого веб-документов.

Сформированный список ресурсов был помещен в массив json, где элементами являются категории сайтов, в каждой категории есть свой список сайтов (рис. 1-2).

Рис. 1. Фрагмент входного файла.

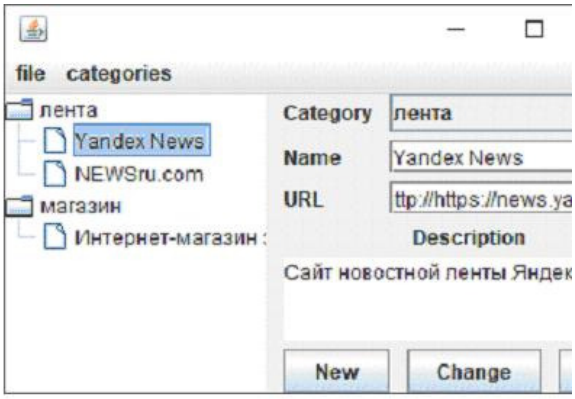

Рис. 2. Интерфейс программы для работы с файлом входного списка сайтов.

В результате изучения структуры сайтов были выделены основные теги, которые были объединены в маркеры. Как и список тестового набора, маркеры хранились в файлах в формате json, так как на данном этапе необходимости в базе данных нет. В одном маркере хранится информация о возможном названии блока информации, тегах и структуре дочерних элементов. Таким образом у маркеров иерархическая структура для некоторых типов сайтов, например ленты новостей или интернет-магазина. Например, проверка того, что ресурс является списком новостей в алгоритме упрощенно выглядит следующим образом:

1. Поиск элемента с текстом «все новости», «лента новостей», «последние новости» и т.д. Количество поисковых фраз на этом этапе зависит от размера словаря. Если элемент был обнаружен, алгоритм переходит к следующему этапу, в противном случае проверяет страницу на маркеры другого типа. Если проверки по всем маркерам дали отрицательный результат, то сайт помечается маркером «неизвестный» и алгоритм переходит к пункту четыре;

2. На втором этапе выявляется список новостей. Для этого проверяются дочерние элементы текущего элемента, если их нет, то алгоритм переходит по иерархии тегов веб-документа вверх до элемента, который включает текущий элемент и другие элемент с текстовым содержимым;

3. Алгоритм в цикле проходит по каждому элементу и проверяет его на соответствие маркерам «краткого описания». Кроме списка новостей с коротким описанием данный алгоритм сможет определить список статей или записей в блоге, так как подобные ресурсы имеют схожую структуру;

4. Загружается 5 дополнительных страниц ресурса для определения статического содержимого страницы: шапка, меню, нижняя панель и т.д. Для более точного определения количество страниц можно увеличить, но в рамках текущего эксперимента 5 страниц было достаточно. Если на первом этапе не было определено основное содержимого страницы, то на этом этапе все, что не является статическим содержимым, помечается маркером основного содержимого;

5. Выявление рекламного содержимого путем сравнения HTML и JavaScript кода с тем, что

предоставляют рекламные компании Google, Яндекса и одна из рекламных сетей Darknet. В дальнейшем набор маркеров для рекламного содержимого был расширен;

6. Выделение содержимого тега <head>, а именно мета-тегов. На этом этапе извлекались только те мета-теги, которые содержали текст, так как в <head> достаточно часто храниться ссылка на библиотеки или фреймворки JavaScript, а также стили CSS;

7. Выделение картинок и внешних ссылок на странице.

Таким образом на выходе программного модуля получается размеченная HTML страница или ее фрагменты, в зависимости от установленных в модуле настроек.

V. Результаты

В тестовый набор было включено 50 ресурсов, среди которых 28 ресурсов сети Darknet. В программный модуль было отправлено по одной странице от каждого ресурса, то есть именно та страница, которая была во входном файле. Результаты работы программного модуля представлены в таблице ниже.

Таблица 1. Результаты работы модуля.

Количество ресурсов	Процент маркировки	Выявлено основное содержимое
16	95-100%	+
15	90-95%	+
11	80-90%	+
5	50-80%	+ / -
3	меньше 50%	-

В большинстве случаев модуль успешно обнаруживал основное содержимое страницы. В некоторых случаях модуль допускал ошибки, разделяя текст одной статьи на отдельные элементы, скорее всего из-за особенностей структуры веб-страниц. Модуль сохранял страницы с маркерами в html формате, что давало возможность открыть их с помощью браузера и проверить, насколько успешно была размечена страница. Ниже на рисунке 3 представлена страница одного из ресурсов тестового набора до маркировки, а на рисунке 4 после маркировки.

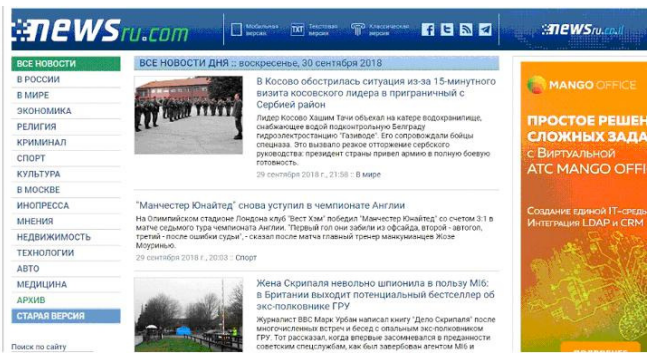

Рис. 3. Веб-страница до маркировки

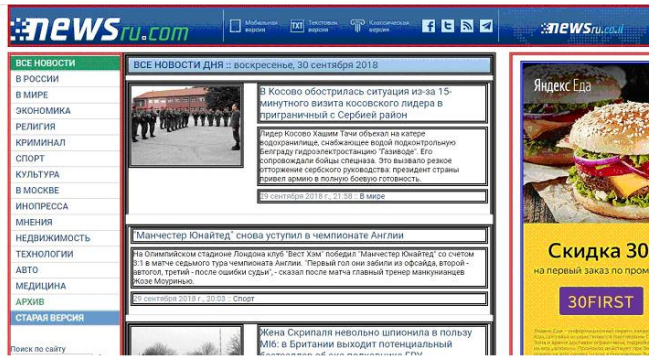

Рис. 4. Веб-страница после маркировки.

Разработанный модуль был совмещен с разработанным во время проведения предыдущего исследования [2] программным комплексом. В результате программный комплекс из предыдущего исследования стал допускать на 10% меньше ошибок при принятии решения о наличии на ресурсе запрещенного содержимого.

VI. Заключение

На основании проведенного анализа и результатов эксперимента был сделан ряд выводов, в том числе дальнейшие направления исследований.

По информации, представленной в разделе «Анализ проблемы» можно сделать следующие выводы:

1. Количество запрещенного контента продолжает расти, не смотря на активную борьбы с ним со стороны государственных органов. Эти сведения подтверждаются отчетами, представленными на официальном сайте IWF. Если посмотреть на динамику роста запрещенного контента в странах за последние несколько лет, то становится понятно, что злоумышленники перемещают свои ресурсы из одной страны в другую и чаще всего при распространении запрещенного содержимого для определенной страны используют пространство адресов другой страны. Таким образом правительство не может «попросить» владельца хостинга об удалении ресурса, распространяющего запрещенный контент, так как он находится за пределами юрисдикции данной страны и регулируется другим законодательством;

2. О наличии запрещенного содержимого на конкретном ресурсе или его отдельной странице организации узнают по жалобам, которые приходят от пользователей сети Интернет, а также в результате внутренних проверок ресурсов сети. Проверка наличия запрещенного содержимого на том или ином ресурсе осуществляется экспертами, то есть веб-документы проверяются преимущественно в ручном режиме, а значит нет программных средств для автоматической проверки, что значительно могло бы ускорить процесс проверки огромного количества веб-страниц;

3. Несмотря на обмен между странами информацией о наличии запрещенного контента на ресурсах, например, с помощью общеевропейской горячей линии INHOPE , злоумышленникам удается избежать ответственности за создание запрещенного ресурса или распространение

запрещенной информации.

В результате проведенного эксперимента был разработан программный модуль, позволяющий разделять содержимого веб-документа на отдельные фрагменты. Таким образом программный модуль для одной страницы создает множество отдельных частей, которые можно проверить и совместить результаты проверки этих частей. В дальнейших исследованиях будет продолжено исследование проблемы распространения запрещенного содержимого и разрабатываться программный комплекс по принятию решения о наличии на странице запрещенного содержимого.

Библиография

[1] Фролов. А.А., Сильнов Д.С. Исследование механизмов распространения запрещенного содержимого в DarkNet // Современные информационные технологии и ИТ-образование – 2017. – 13, №4. – С. 216-224.

[2] Frolov A., Silnov D., Geraschencko Y., Sadretdinov A., Kiamov A. Research of mechanisms counteracting the distribution of prohibited content on the Internet // 2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering // Moscow, 2018.

[3] Ельчанинова Н.Б. Проблемы совершенствования законодательства в сфере ограничения доступа к противоправной информации в сети Интернет // Общество: политика, экономика, право – 2017. – №12. – С. 119-121.

[4] Марценюк А.Г. Запрещенная информация и ее место в системе информационных отношений // Гражданин и право – 2018. – №5. – С. 62-62.

[5] Романова Л.И., Дремлюга Р.И. Преступный мир Darknet // Юридическая наука и практика – 2018. – 14. - №1. – С. 52-60.

[6] Dr. Digvijaysinh Rathod. Darknet Forensics // International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) // 2017. Vol. 6. PP. 77-79.

[7] Alin Teodorus Drăgan. Child pornography and child abuse in cyberspace // Journal of legal studies // 2018. Vol. 21, № 35. PP. 52-60.

[8] Олейникова Е.А. Противодействие экстремизму в сети Интернет // Законность – 2016. – №5(979). – С. 6-9

[9] Валеев А.Х. Противодействие экстремизму в сети Интернет // Труды академии МВД республики Таджикистан – 2015. – №3. – С. 55-56.

[10] Бижоева М.К. Противодействие экстремизму в сети Интернет // Ростовский научный журнал – 2018. – №1. – С. 43-48.

[11] Рыдченко К.Д. Запрещенная для детей информация: легальная трактовка и толкование правоприменителя // Информационное право – 2014. - №5. – С. 16-21.

[12] Kawade, Kavita Oza. Content-based SMS spam filtering using machine learning technique // International Journal of Computer Engineering and Applications // 2018. Vol. 7, № 4. PP. 625-630.

[13] Bhavish Khanna Narayanan, M. Rajasekharababu, J. Moses SharonMoses. Adult content filtering: Restricting minor audience from accessing inappropriate internet content // Education and Information Technologies // 2018

[14] Prof. G. N. Purohit. Dr. Priti Singh. Praveen Dangi. Content Filtering on Social Networking Sites with Fuzzy Logic // International Journals of Advanced Research in Computer Science and Software Engineering // 2017. Vol. 7, № 6. PP. 175-179.

[15] Carl-Maria Morch, Louis-Philippe Cote, Laurent Corthesy-Blondin. The Darknet and suicide // Journal of Affective Disorders // 2018. Vol. 241. PP. 127–132

[16] Yannikos York, Schäfer Annika, Steinebach Martin. Monitoring Product Sales in Darknet Shops // International Conference on Availability, Reliability and Security // 2018

[17] Негодин М.Ю. Ограничение доступа к запрещённой информации в сети Интернет // Участие студентов в обеспечении комплексной безопасности, профилактике экстремизма, терроризма и антикоррупционной деятельности в учебных заведениях – 2017. – С.69-70.

[18] Щурова А.С. Об ограничении доступа к сайтам в сети «интернет», содержащим запрещённую информацию о наркотических средствах // Антинаркотическая безопасность – 2016. – №1(6). – С. 29-31.

[19] Кутовой Н.Н., Романова Е.А. Актуальные проблемы в работе системы Единого реестра запрещённых сайтов // Вестник молодых ученых и специалистов Самарского государственного университета – 2016. – №2(9). – С. 102-106.

[20] Кутовой Н.Н., Романова Е.А. Исследование проблем деятельности Единого реестра запрещенных сайтов // Научное сообщество студентов – 2018 – С. 261-263.

[21] Кутовой Н.Н., Романова Е.А. Анализ Системы Единого реестра запрещенных сайтов // Наука, образование, общество: тенденции и перспективы развития – 2018. – С. 162-163.

[22] Балашов А.Н., Правовое регулирование Интернет-отношений: основные проблемы и практика реализации в России // Среднерусский вестник общественных наук – 2016. – 11, №2 – С. 113-118.

[23] Кузьмин А.Е., Кульназарова А.В. Анализ практики ограничения доступа к контенту в Рунете со стороны органов государственной власти за 2014–2017 // Вестник Омского университета. Серия: Исторические науки – 2017. – №3(15). – С. 429-434.

[24] Jie Chen, Jun Zhou, Kun Pan. The Security of Key Derivation Functions in WINRAR // Journal of computers // 2013. Vol. 8, № 9. PP. 2262-2268.

[25] Meiqi Wang, Xuebin Wang, Jinqiao Shi. Who are in the Darknet? Measurement and Analysis of Darknet Person Attributes // 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC) // 2018. PP. 948-955.

[26] Сидорова Е.А., Кононенко И.С., Загорулько Ю.А., Подход к фильтрации запрещенного контента в веб-пространстве // Аналитика и управление данными в областях с интенсивным использованием данных – 2017. – С. 94-101.

[27] Сидорова Е.А., Боровикова О.И. Подход к жанровой классификации текстовых ресурсов // Информационные технологии и системы – 2017. – С. 264-269

Analysis of mechanisms for detection prohibited content on the Internet

A.A. Frolov, D.S. Silnov, A.M. Sadretdinov

Abstract— **In this article was researched such an issue as prohibited content spreading in the Internet network and a research of the mechanisms for detecting this content on web pages. An analysis is made of existing solutions in this area, including legislative measures that ensure the process of removing prohibited content from the resource or restricting access to it, information from reports of organizations that restrict access to prohibited resources or individual pages of these resources, which indicates an increase in the total number of prohibited contents on the network. The results of the analysis of existing methods for detecting prohibited content are presented, on the basis of which decisions are made on the need to develop new approaches to the detection of this content. Great attention is paid to the problem of protecting children on the Internet, as well as the problem of distributing prohibited content through hidden networks and other threats to network users arising from these networks. We describe an experiment that was conducted to analyze the content of web resources in order to improve the quality of detection mechanisms for prohibited content on the Internet and the Darknet network. During the experiment, a software module was developed to extract and process Web pages, which made markup of the web page with special markers to separate the content of the web document into groups. The data obtained from the results of the experiment contain the results of testing the developed module and examples of its operation. The final part of the article contains a set of methods that the author intends to use in future studies, in which it is planned to develop a software package for making decisions about the presence of prohibited content on the Internet web resources.**

Keywords— **Internet; filtration; prohibited content.**

REFERENCES

[1] Frolov. A.A., Sil'nov D.S. Issledovanie mehanizmov rasprostranenija zapreshhennogo soderzhimogo v DarkNet // Sovremennye informacionnye tehnologii i IT-obrazovanie – 2017. – 13, #4. – S. 216-224.

[2] Frolov A., Silnov D., Geraschencko Y., Sadretdinov A., Kiamov A. Research of mechanisms counteracting the distribution of prohibited content on the Internet // 2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering // Moscow, 2018.

[3] El'chaninova N.B. Problemy sovershenstvovanija zakonodatel'stva v sfere ogranichenija dostupa k protivopravnoj informacii v seti Internet // Obshhestvo: politika, jekonomika, pravo – 2017. – #12. – S. 119-121.

[4] Marcenjuk A.G. Zapreshhennaja informacija i ee mesto v sisteme informacionnyh otnoshenij // Grazhdanin i pravo – 2018. – #5. – S. 62-62.

[5] Romanova L.I., Dremljuga R.I. Prestupnyj mir Darknet // Juridicheskaja nauka i praktika – 2018. – 14. - #1. – S. 52-60.

[6] Dr. Digvijaysinh Rathod. Darknet Forensics // International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) // 2017. Vol. 6. PP. 77-79.

[7] Alin Teodorus Drăgan. Child pornography and child abuse in cyberspace // Journal of legal studies // 2018. Vol. 21, # 35. PP. 52-60.

[8] Olejnikova E.A. Protivodejstvie jekstremizmu v seti Internet // Zakonnost' – 2016. – #5(979). – S. 6-9

[9] Valeev A.H. Protivodejstvie jekstremizmu v seti Internet // Trudy akademii MVD respubliki Tadzhikistan – 2015. – #3. – S. 55-56.

[10] Bizhoeva M.K. Protivodejstvie jekstremizmu v seti Internet // Rostovskij nauchnyj zhurnal – 2018. – #1. – S. 43-48.

[11] Rydchenko K.D. Zapreshhennaja dlja detej informacija: legal'naja traktovka i tolkovanie pravoprimenitelja // Informacionnoe pravo – 2014. - #5. – S. 16-21.

[12] Kawade, Kavita Oza. Content-based SMS spam filtering using machine learning technique // International Journal of Computer Engineering and Applications // 2018. Vol. 7, # 4. PP. 625-630.

[13] Bhavish Khanna Narayanan, M. Rajasekharababu, J. Moses SharonMoses. Adult content filtering: Restricting minor audience from accessing inappropriate internet content // Education and Information Technologies // 2018

[14] Prof. G. N. Purohit. Dr. Priti Singh. Praveen Dangi. Content Filtering on Social Networking Sites with Fuzzy Logic // International Journals of Advanced Research in Computer Science and Software Engineering // 2017. Vol. 7, # 6. PP. 175-179.

[15] Carl-Maria Morch, Louis-Philippe Cote, Laurent Corthesy-Blondin. The Darknet and suicide // Journal of Affective Disorders // 2018. Vol. 241. PP. 127–132

[16] Yannikos York, Schäfer Annika, Steinebach Martin. Monitoring Product Sales in Darknet Shops // International Conference on Availability, Reliability and Security // 2018

[17] Negodin M.Ju. Ogranichenie dostupa k zapreshhjonnoj informacii v seti Internet // Uchastie studentov v obespechenii kompleksnoj bezopasnosti, profilaktike jekstremizma, terrorizma i antikorrupcionnoj dejatel'nosti v uchebnyh zavedenijah – 2017. – C.69-70.

[18] Shhurova A.S. Ob ogranichenii dostupa k sajtam v seti «internet», soderzhashhim zapreshhjonnuju informaciju o narkoticheskih sredstvah // Antinarkoticheskaja bezopasnost' – 2016. – #1(6). – S. 29-31.

[19] Kutovoj N.N., Romanova E.A. Aktual'nye problemy v rabote sistemy Edinogo reestra zapreshhjonnyh sajtov // Vestnik molodyh uchenyh i specialistov Samarskogo gosudarstvennogo universiteta – 2016. – #2(9). – S. 102-106.

[20] Kutovoj N.N., Romanova E.A. Issledovanie problem dejatel'nosti Edinogo reestra zapreshhennyh sajtov // Nauchnoe soobshhestvo studentov – 2018 – S. 261-263.

[21] Kutovoj N.N., Romanova E.A. Analiz Sistemy Edinogo reestra zapreshhennyh sajtov // Nauka, obrazovanie, obshhestvo: tendencii i perspektivy razvitija – 2018. – C. 162-163.

[22] Balashov A.N., Pravovoe regulirovanie Internet-otnoshenij: osnovnye problemy i praktika realizacii v Rossii // Srednerusskij vestnik obshhestvennyh nauk – 2016. – 11, #2 – S. 113-118.

[23] Kuz'min A.E., Kul'nazarova A.V. Analiz praktiki ogranichenija dostupa k kontentu v Runete so storony organov gosudarstvennoj vlasti za 2014–2017 // Vestnik Omskogo universiteta. Serija: Istoricheskie nauki – 2017. – #3(15). – S. 429-434.

[24] Jie Chen, Jun Zhou, Kun Pan. The Security of Key Derivation Functions in WINRAR // Journal of computers // 2013. Vol. 8, # 9. PP. 2262-2268.

[25] Meiqi Wang, Xuebin Wang, Jinqiao Shi. Who are in the Darknet? Measurement and Analysis of Darknet Person Attributes // 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC) // 2018. PP. 948-955.

[26] Sidorova E.A., Kononenko I.S., Zagorul'ko Ju.A., Podhod k fil'tracii zapreshhennogo kontenta v veb-prostranstve // Analitika i upravlenie dannymi v oblastjah s intensivnym ispol'zovaniem dannyh – 2017. – S. 94-101.

[27] Sidorova E.A., Borovikova O.I. Podhod k zhanrovoj klassifikacii tekstovyh resursov // Informacionnye tehnologii i sistemy – 2017. – S. 264-269