Я. Д. Хлыстикова, А. А. Оленников
Тюменский государственный университет, г. Тюмень
УДК 004.422

ПРОГРАММНОЕ РЕШЕНИЕ ДЛЯ ПОИСКА И АНАЛИЗА ЗАПРЕЩЕННОЙ ИНФОРМАЦИИ В ГЛОБАЛЬНОЙ СЕТИ

Аннотация. В работе представлено программное решение для поиска и подробного анализа запрещенного контента в глобальной сети на территории Российской Федерации, позволяющее выявлять географию и другие технические данные серверов, распространяющие запрещенную информацию.

Ключевые слова: информация, глобальная сеть, программное решение, анализ, запрещенная информация, анализ.

Введение. С каждым днем в глобальной сети появляется множество веб-ресурсов доступные пользователям и содержащие информацию, распространение которой в Российской Федерации запрещено. Тематики этих ресурсов связаны с призывом межнациональной вражды, и распространение запрещенных препаратов, насилие, мошенничество, да и многое другое. Проблема своевременного выявления таких ресурсов стоит очень остро [1]. Последствия от вовремя необнаруженных веб ресурсов могут быть плачевными [2].

В настоящее время существуют решения, которые позволяют выполнить анализ контента в глобальной сети, однако эти ресурсы работают только на верхних уровнях и в основном недоступны экспертам занимающиеся киберрасследованиями или региональным провайдерам. Более того, такие решения не имеют возможности прогнозирования ущерба от таких ресурсов.

В этой работе пока положено начало программного решения, которое позволяет анализировать ресурсы глобальной сети, а критерии оценивания ущерба будут освещены в отдельных трудах.

Проблема исследования. В эпоху «Интернета» не только существует много возможностей поиска информации, но и связанные с проблемой ограничения информации. Каждую минуту или секунду появляется новый сайт, создаются программы, развиваются новые технологии. Не всегда можно анализировать тот объем новой информации, который попадает в свободный доступ. Распространение запрещенной информации в Российской Федерации карается уголовной, административной ответственностью [3]. Также не желательная информация может навредить морально человеку. Существуют частичное решение этой проблемы такие как список запрещенных доменов. Программный, которые частично выполняют функции поиска по запросу с различным функционалом анализа и глубинного поиска [4, 5]. Все эти решения не рассчитывают коэффициент ущерба от такой информации.

Материалы и методы. Узнать ip адрес или домен можно разными способами: использовать стандартные команды такие как ping, nslookup, с помощью внешних специализированных ресурсов таких как google dig, сайт 2ip. Первый вариант не подходит из-за того, что нужно запускать фоновый процесс, а он, может быть, не один соответственно компьютер может перейти в аварийное состояние, благодаря перегрузке. Также стоит отметить, что этот способ довольно неинформативный для решения задач разрабатываемой программы. Второй вариант в свою очередь является плюсом так как по мимо нужной информации можно узнать

дополнительную по задаваемым ip или доменам. Большинство внешних сервисов работаю с api ключами, которые в свою очередь платные или предлагаются на пробный период использования, что является не очень удобным для будущего программного решения.

При поиске ip адреса и домена существует два метода. Первый метод составление собственной базы данных ip и доменов. У этого метода есть большой недостаток. С каждой минутой или секундой создается как минимум один сайт соответственно за этот короткий промежуток времени базу данных нужно обновлять из этого следует очень трудоемкий процесс и затрачивается очень много времени. Также есть положительная сторона того, что этот вариант не зависит от внешних ресурсов.

Второй метод использовать существующие решения, у которых есть своя актуальная база данных ip адресов и доменов. Также есть недостаток, зависимость от ресурса. Если с ресурсом что-то случится, то не будет доступа к базе данных с ip и доменами. Чтобы решить эту проблему в разработке используется два ресурса, один основной, другой запасной и вероятность того, что неисправный одновременно будут сразу оба очень мала.

В работе совмещены оба метода. Работает это таким образом программа по запросу обращается на внешний ресурс, с него берет нужную информацию, а именно ip адрес и домен. Далее эту информацию сохраняет в свою базу данных, таким образом сама база данных пополняется актуальной информацией доменов и ip. Также разрабатываемая программа работает уже со своей базой данных доменов и ip.

Разрабатываемая программа осуществляет поиск с изначальными данными ip или доменом. Все домены, лежащие на одном ip адресе анализируются, соответственно и наоборот все ip на одном домене также проходит анализ. Программа анализирует ссылки на домене или ip, который проверяет. Сайты с текстовой и графической информацией в том числе. Алгоритм представлен на рис. 1.

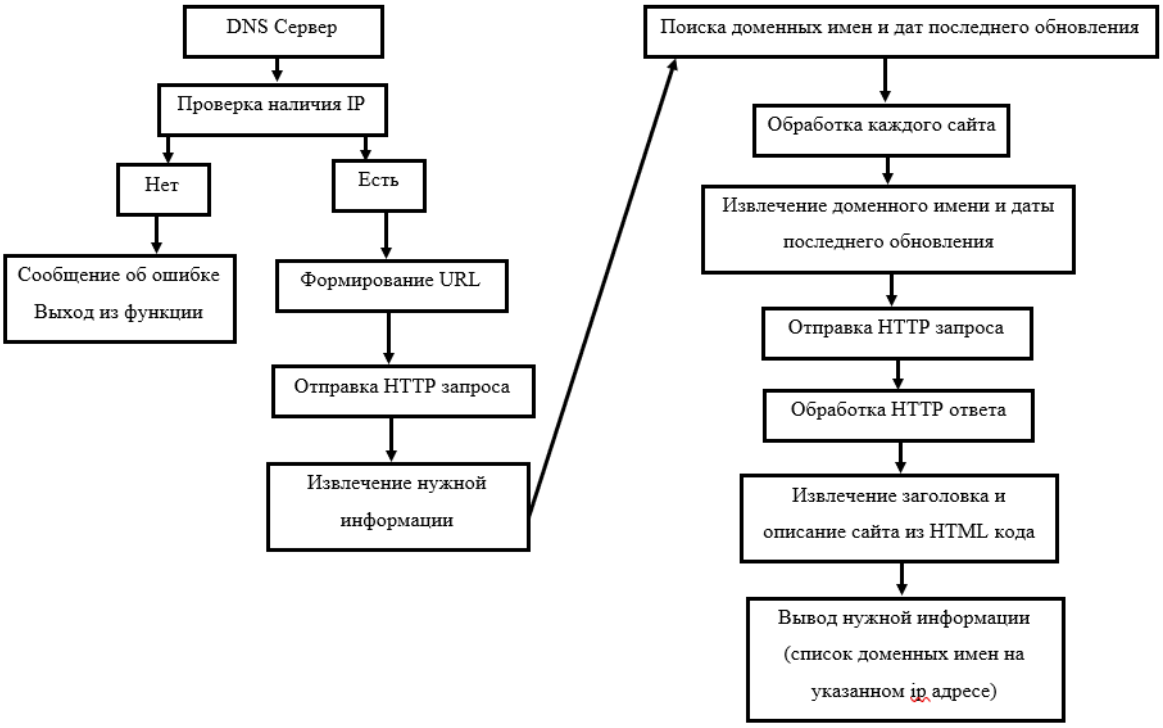

Рис. 1. Алгоритм поиска доменных имен на ip адресе

Метод поиска реализован таким образом, что из исходных данных получаем домен или ip. Если это был домен, программа переходит по этому домену и далее все ссылки с указанием этого домена также анализируются программой. Если это был ip адрес, то находящейся домен по указанному ip также анализируются точно таким же образом. Ничем не отличается анализ при нескольких ip или доменов, проверяется каждый ip и домен отдельно.

Стоит отметить, что вся информация находящейся на изначальных данных, которые проходят анализ сохраняется в текстовый файл и базу данных, что при дальнейшей работе программы помогает ускорить работу поиска и анализа, также сохраннее информации поможет формировать доказательную базу, например в расследовании.

В анализе текстовой информации большую помощь приносит парсинг страниц [6, 7]. Парсинг помогает собрать нужную информацию и удалить не нужную. Алгоритм парсинга представлен на рис. 2.

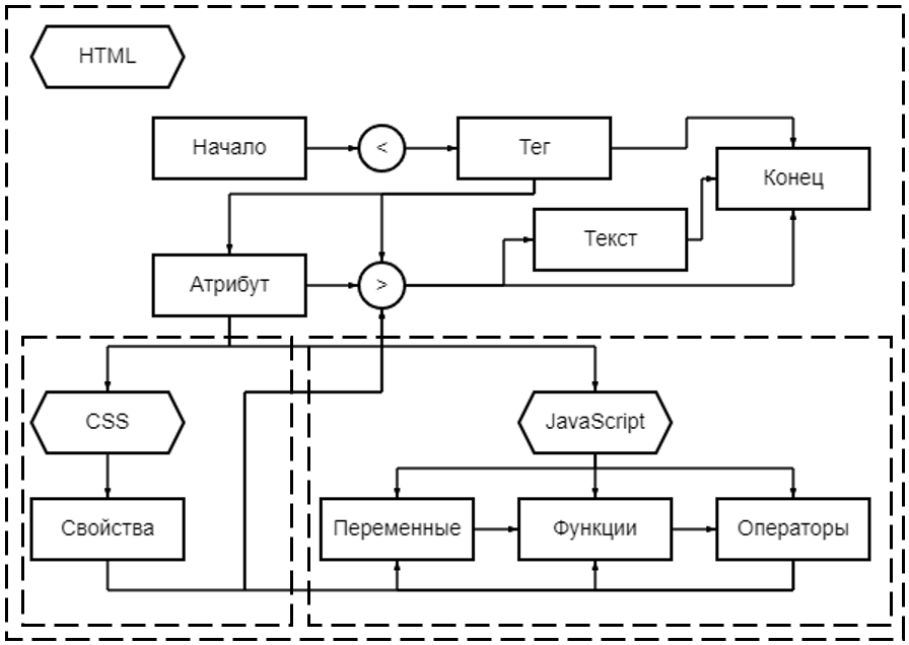

Рис. 2. Алгоритм парсинга страницы

В разрабатываемой программе реализован симбиоз нескольких методов, а реализовано это таким образом, что в базе данных хранится словарь фраз и слов, с которым сравнивается вся найденная текстовая информация в свою очередь вся текстовая информация, с которой работает программа разбивается на предложения. При нахождении совпадений сохраняется предложение с найденной фразой или словом при этом идет подсчет количества слов найденных, общее количество слов, а также количество слов в предложении [8].

Метод анализа графических изображений в разрабатываемой программе основан на нейронной сети, обученной на своем датасете из разного рода картинок, связанных с запрещенной информацией.

В работе используется модель сверточной нейронной сети. Основа работы, которой лежит в операции свертки над парой матриц по следующей формуле:

$$C = A \times B \text{ размера } (n_x - m_x + 1) \times (n_y - m_y + 1),$$

где A размера $n_x \times n_y$ и B размера $m_x \times m_y$.

Обучение модели описывается переходом от особенностей изображений к абстрактным представлениям, далее еще к более абстрактным и так далее до представлений высокого уровня. При этом в основе обучения модель отбрасывает маловажные части изображения и фокусируется на более существенных, алгоритм представлен на рис. 3.

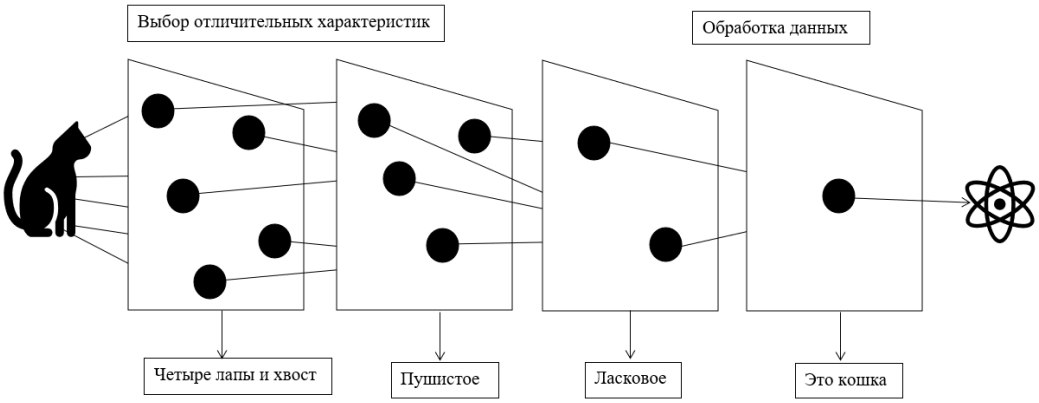

Рис. 3. Обобщенный пример работы сверточной нейронной сети

Расчет коэффициента осуществляется на основе анализа всей собранной текстовой информацией и графической информацией с веб ресурсов.

Результат. Итог работы программного решения — разработанный сервис, позволяющий анализировать контент на основе взятого домена, ip адреса или диапазона ip адресов. Результатом анализа веб ресурсов является три вида отчетов по всей собранной информации, отражающие географию и другие технические данные о серверах, транслирующие запрещенных контент в глобальную сеть, также проведенного анализа запрещенной информации [9, 10]. Фрагмент графического интерфейса представлен на рис. 4.

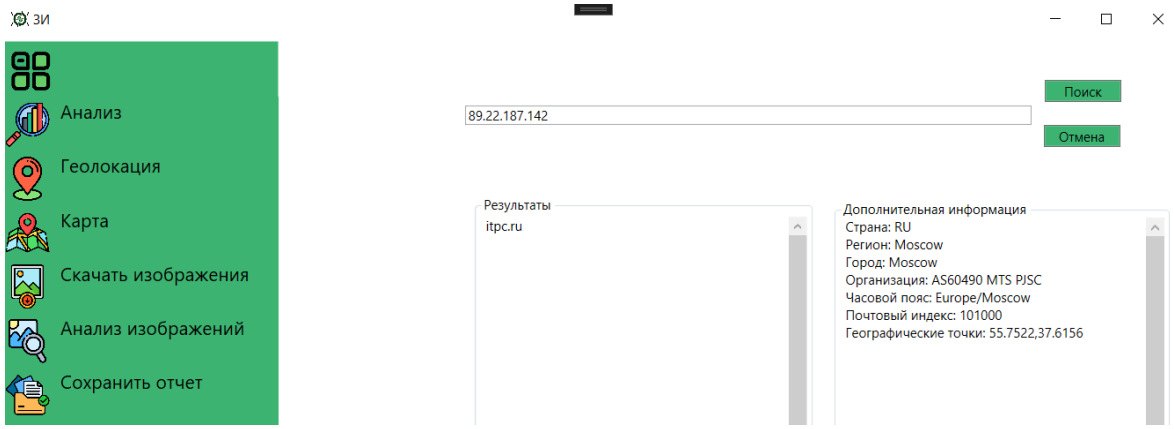

Рис. 4. Интерфейс программы

Заключение. Предложенное программное решение уже позволяет выявлять запрещенный контент, что уже становится очень полезным для экспертов и лиц занимающихся киберрасследованиями. В настоящее время на кафедре Информационной безопасности ТюмГУ ведется активное работа по развитию данного сервиса. Планируется ввести критерии оценивания ущерба от публикуемой запрещенной информации и другая полезная информация, которая будет изложена в отдельных трудах.

СПИСОК ЛИТЕРАТУРЫ

1. Исследование механизмов распространения запрещенного содержимого в Darknet / А.А. Фролов, Д.С. Сильнов // Современные информационные технологии и ИТ-образование. — 2017. — Т. 13. — № 4. — С. 216-224.
2. Проблемы отрицательного влияния интернета на нравственное воспитание подростков в информационном пространстве и пути решения / Э.И. Атагимова // Правовая информатика. — 2013. — № 1. — С. 21-24.
3. Федеральный закон «Об информации, информационных технологиях и о защите информации» от 27.07.2006 № 149-ФЗ. — URL: https://eais.rkn.gov.ru/docs/149.pdf (дата обращения 30.10.2023).
4. Брумштейн Ю.М., Бондарев А.А., Кузьмина А.Б. Компьютерное обеспечение и вычислительная техника // Научно-техническая информация. 2014. — № 4. — С. 40-54.
5. Суслов А.В., Ажмухвмедов И.М. Программное обеспечение для выявления запрещенного текстового контента // Научно-техническая информация. — 2018. — № 1 (41). — С. 185-195.
6. Антошин П.И., Каплин В.И. Автоматический анализ текстов. Синтаксический и семантический анализ // Научно-техническая информация. — 2017. — № 6. — С. 211-213.
7. Никитин А.В. Анализ методов синтаксического анализа сайтов средствами технологической платформы «1С: предприятие» // Научно-техническая информация. Том 1. — 2016. — С. 110-113.
8. Смирнов И.В., Шелеманов А.О. Семантико-синтаксический анализ естественных языков. Часть I. Обзор методов синтаксического и семантического анализа текстов // Научно-техническая информация. — 2013. — № 1. — С. 41-54.
9. Mydocx.ru Что такое поиск информации в сети. — 2016. — URL: https://mydocx.ru/11-86672.html (дата обращения 01.11.2023).
10. Фролов А.А., Сильнов Д.С., Садретдинов А.М. Анализ механизмов обнаружения запрещенного содержимого в сети Интернет // Научно-техническая информация. — 2019. — Vol. 7, no 1. — С. 90-96.