Проблемы поиска противоправной информации в сети «Интернет»

С.В. Волошин
Центр информационно-аналитических систем
АУ «Югорский НИИ
информационных технологий»
Ханты-Мансийск, Россия
e-mail: VoloshinSV@uriit.ru

Аннотация

В настоящее время одной из актуальных задач информационного поиска является выявление в сети "Интернет" противоправной информации. По данным Роскомнадзора, за третий квартал 2017 года количество обращений от граждан увеличилось в 3.5 раза, по сравнению с аналогичным периодом прошлого года. В докладе приведены некоторые особенности поиска противоправной информации и описан ряд проблем, возникающих при ручном поиске информации. Предложены пути решения перечисленных проблем. Приведено описание модели сбора и выдачи информации для оптимизации выявления противоправной информации экспертами.

1. Введение

В сети «Интернет» существует множество ресурсов, распространяющих противоправную информацию, выявление которой является важной задачей [1]. Выявление этих ресурсов в сети осуществляется экспертами, владеющими специализированными знаниями в области выявления противоправной информации. Как правило, эксперты осуществляют поиск в «ручном режиме», без использования средств автоматизации. Однако такой подход к поиску и выявлению противоправной информации без использования средств автоматизации имеет ряд недостатков, существенно снижающих общую эффективность работы группы экспертов.

В данной работе будет рассмотрена оптимизация процесса выявления материалов с точки зрения автоматизации. Также будет предложена общая модель сбора и выдачи информации экспертам.

Интерес представляет оригинальный метод поиска противоправной информации с учетом специфики как анализируемых данных, так и работы группы экспертов.

В настоящее время существует ряд информационных систем, осуществляющих специализированный поиск в различных предметных областях [2]. Но, в большинстве случаев, информация об их структуре и применяемых методах обработки данных не раскрывается. Зачастую они предоставляются по технологии SaaS (англ. Software as a Service), что неприемлемо с учетом специфики обрабатываемых данных.

2. Проблемы и пути решения

2.1. Виды информации и направления поиска

В сети «Интернет» можно выделить ряд категорий противоправной информации, в том числе:

- информация о способах, методах разработки, изготовления и использования наркотических средств, психотропных веществ и их прекурсоров (п.1 ч.5 ст.15.1 Федерального закона от 27.07.2006 № 149-ФЗ)
- информация о способах совершения самоубийства, а также призывов к совершению самоубийства (п.1 ч.5 ст.15.1 Федеральный закон от 27.07.2006 № 149-ФЗ)
- информация, содержащая призывы к массовым беспорядкам, осуществлению экстремистской деятельности, участию в массовых (публичных) мероприятиях, проводимых с нарушением установленного порядка (ст.15.3 Федерального закона от 27.07.2006 № 149-ФЗ)
- информация, содержащая публичные призывы к осуществлению террористической деятельности или публично оправдывающая терроризм (ст.10.4 Федерального закона от 27.07.2006 № 149-ФЗ)

С другой стороны, с точки зрения структуры содержимого материалов, можно выделить следующие виды информации:

- текстовая,
- графическая,
- аудиозаписи,
- видеозаписи.

Обработка каждой категории противоправной информации и каждого вида содержимого материалов имеет свою специфику, которая

учитывается при организации сбора, хранения и обработки информации [3].

2.2. Постановка задачи

Для определения проблем, которые требуют решения, обозначим типовую последовательность действий эксперта при выявлении материалов, содержащих противоправную информацию. Последовательность в общем виде выглядит следующим образом:

1. Эксперт формирует запрос, по которому ожидает получить список информационных ресурсов, содержащих ключевые слова, характерные для противоправной информации.
2. Эксперт направляет запрос в поисковые системы. От поисковых систем получает список проиндексированных ресурсов, которые содержат слова, характерные для противоправной информации.
3. Проверяет эти ресурсы на наличие противоправной информации. В случае если она есть – отправляют идентификатор ресурса в Роскомнадзор, который выступает в качестве контролирующего органа.

При таком подходе возникают следующие проблемы:

1. Устаревание поисковых запросов: для того, чтобы поисковые запросы пользователей были актуальны, экспертам необходимо тратить время на выявление новых ключевых понятий, направлений в предметной области и составление поисковых запросов по ним.
2. Повторяемость проверки материалов: эксперты работают независимо друг от друга, вследствие чего одна и та же информация может быть проверена множество раз. Такой процесс неэффективен и может приводить к противоречивой оценке. Прогнозировать эффективность работы экспертов в этом случае становится весьма затруднительным, так как трудозатраты на один материал (вследствие неопределенного числа проверок) могут значительно различаться.
3. Отсутствие верификации отчетов экспертов: пользователи ведут журнал для отчетности в ручном режиме, и нет никакого способа убедиться, что записи журнала не подделаны или искажены (случайно или намеренно).
4. Малое количество предварительных данных для экспертизы материалов: пользователи для экспертизы выбирают материалы из списка, руководствуясь минимумом информации о материале (пользователи при работе с большинством популярных поисковых систем используют только адрес веб-страницы, домен и короткий текстовый фрагмент, выдаваемый поисковой системой).

2.3. Централизация

В качестве основы для решения вышеуказанных проблем предлагается разработать автоматизированную информационную систему поиска и анализа информации в сети «Интернет» [4].

Отметим, что централизация только частично решает проблему устаревания поисковых запросов. Последняя проблема из списка в предыдущем пункте не столь тривиальна и будет рассмотрена отдельно.

Применительно к проблеме отсутствия верификации отчетов: централизованная система позволяет журналировать любую работу экспертов (как и результат этой работы), что, при отсутствии у пользователя прав на изменение журнала, обеспечивает практическую невозможность подмены или искажения данных в отчете.

Проблема повторяемости проверки информационных ресурсов решается при помощи персонализированной выдачи системы: можно не выдавать эксперту на проверку те ресурсы, которые уже находятся на рассмотрении.

Также, централизация работы экспертов позволяет сформировать единую базу поисковых сущностей. Под поисковой сущностью понимается совокупность информационных единиц (ключевых слов, стоп-слов, синонимов, жаргонизмов, графических примитивов и т.п.), характеризующих один класс объектов реального мира и позволяющая с высокой степенью достоверности идентифицировать данный объект.

Единая база поисковых сущностей позволит частично решить проблему устаревания поисковых запросов: при нахождении новых ключевых понятий или направлений в предметной области одним из экспертов и актуализации им списка поисковых сущностей, результаты обновленной поисковой выдачи получат все эксперты.

Предлагаемая общая архитектура состоит из:

- подсистемы хранения,
- подсистемы выполнения задач,
- графического интерфейса пользователя.

Подсистема хранения отвечает за хранение информации, проверку целостности данных и является центральным элементом при общении подсистем.

Графический интерфейс пользователя позволяет экспертам работать с информационной системой. Обеспечивает доступ к системе пользователям и позволяет отображать пользователю предварительные данные об информационном ресурсе для последующей экспертизы материалов.

Подсистема выполнения задач отвечает за логику сбора, анализа и обработки материалов в информационной системе:

- загрузка материалов в целях последующего анализа и извлечения предварительных данных для экспертизы;
- поиск потенциально противоправной информации;
- генерацию плановых отчетов и отчетов по требованию;
- анализ полученной информации и извлечение предварительных данных для экспертизы;
- изменение данных в модели поиска информации на основе анализа полученной информации;
- реализации бизнес-процессов, содержащих взаимодействие с внешними информационными системами (например, Роскомнадзором).

2.4. Извлечение предварительных данных для экспертизы

Извлечение предварительных данных призвано решить проблему малого количества предварительных данных для экспертизы материалов экспертом и обеспечить эксперту поддержку в принятии решения о наличии противоправной информации. Предлагается рассмотреть следующее множество возможных задач:

- извлечение фактов,
- обнаружение дубликатов,
- тематическое моделирование,
- извлечение других видов информации,
- классификация на наличие противоправной информации.

2.4.1. Извлечение дополнительных фактов

Извлечение дополнительных фактов из полученной в результате поиска информации может помочь пользователю быстрее и точнее определить наличие противоправной информации. Извлечение фактов может помочь в принятии решения следующим:

- Выявление в материалах других поисковых сущностей, известных системе [5]. Имея информацию о наличии иных поисковых сущностей, эксперт может решить просмотреть материал целиком, а не только тот фрагмент, который был обозначен поисковой системой, как релевантный. К тому же, это полезно при поиске наркотических веществ: как правило, информационные ресурсы фармакологической направленности содержат описание одного препарата на одной веб-странице. Имея множество совпадений по поисковым сущностям, условная вероятность того, что информационный ресурс – это площадка по продаже наркотических веществ, увеличивается.
- Составление цепочек фактов [6] также может обеспечить пользователя дополнительной информацией для принятия решения. Например, цепочка фактов «название наркотика» - «купить» - «цена» говорит о том, что в материале, скорее всего, содержится информация о распространении наркотиков.
- Извлечение персональных данных [7]: в первую очередь, извлечение персональных данных и контактной информации позволяет создать базу личностей, которую можно далее проанализировать и использовать при актуализации поисковых сущностей и запросов, а также при ранжировании выдачи экспертам. В первую очередь, такая информация будет значима при анализе материалов экстремистской и террористической направленности: в определенных случаях, пропаганда ведется от лица определенных личностей, т.н. «духовных лидеров». Создание базы таких личностей позволит находить в полученной информации упоминание этих личностей и повышать рейтинг этой информации в выдаче.

Для выявления фактов можно использовать семантический анализ, контекстно-свободные грамматики, регулярные грамматики, или алгоритмы машинного обучения [8].

2.4.2. Обнаружение дубликатов

Обнаружение дубликатов подразумевает выявление информации, имеющих определенную степень схожести по структуре и содержанию [9]. Такой информацией могут быть, например, зеркала сайтов или повторная публикация информации (например, «репосты» в соцсетях). Материал, у которого среди выявленных схожих материалов преобладают материалы, оцененные как содержащие противоправную информацию, имеет вероятность содержания противоправной информации выше, чем материал, у дубликаты которого не содержат противоправной информации.

Для обнаружения дубликатов возможно использовать, например, алгоритм Megashingles или Long Sent [10].

2.4.3. Тематическое моделирование

Тематическое моделирование текстов позволит автоматически разделить тексты на направления, что, во-первых, позволит пользователям отложить те материалы, в которых эксперты не обладают компетенциями, а во-вторых, позволит анализировать информацию, имея априорные знания о её контексте.

Для тематического моделирования текстов существует множество алгоритмов, в том числе алгоритмы машинного обучения, например:

- метод опорных векторов [11],
- решающие деревья и ансамбли на их основе [12],

- латентное размещение Дирихле [13],
- нейросетевые модели и т.д [14].

2.4.4. Извлечение других видов информации

Задача извлечения других видов информации предполагает, как наиболее очевидный вариант, извлечение текстовой информации из графической [15]. Такой информацией может быть, например, распознанный текст на изображении. Это важно в следующих задачах:

- Выявление экстремистских материалов. Некоторые изображения содержат в себе кодовые лозунги экстремистских группировок. Вариантов написания чисел и слов может быть множество (шрифт, специальные символы, разнесение лозунга по различным частям изображения), что делает задачу классификации изображений в этом контексте не актуальной. В таком случае имеет смысл поставить задачу распознавания текстов.
- Встречаются материалы, содержащие в себе информацию по распространению наркотиков, в которых контактная информация, цена и названия наркотика представлены в графическом виде («встроены» в изображение), что делает классификацию текстов, фактически, неактуальной.

2.4.5. Классификация на наличие противоправной информации

Задача ранжирования выдачи предполагает выдачу экспертам в первую очередь тех материалов, которые наиболее вероятно содержат в себе противоправную информацию. Вероятность содержания противоправной информации определяется как результат комплексной классификации всей информации, полученной по одному указателю (ссылке). Пример возможного комплексного классификатора приведен на рисунке 1:

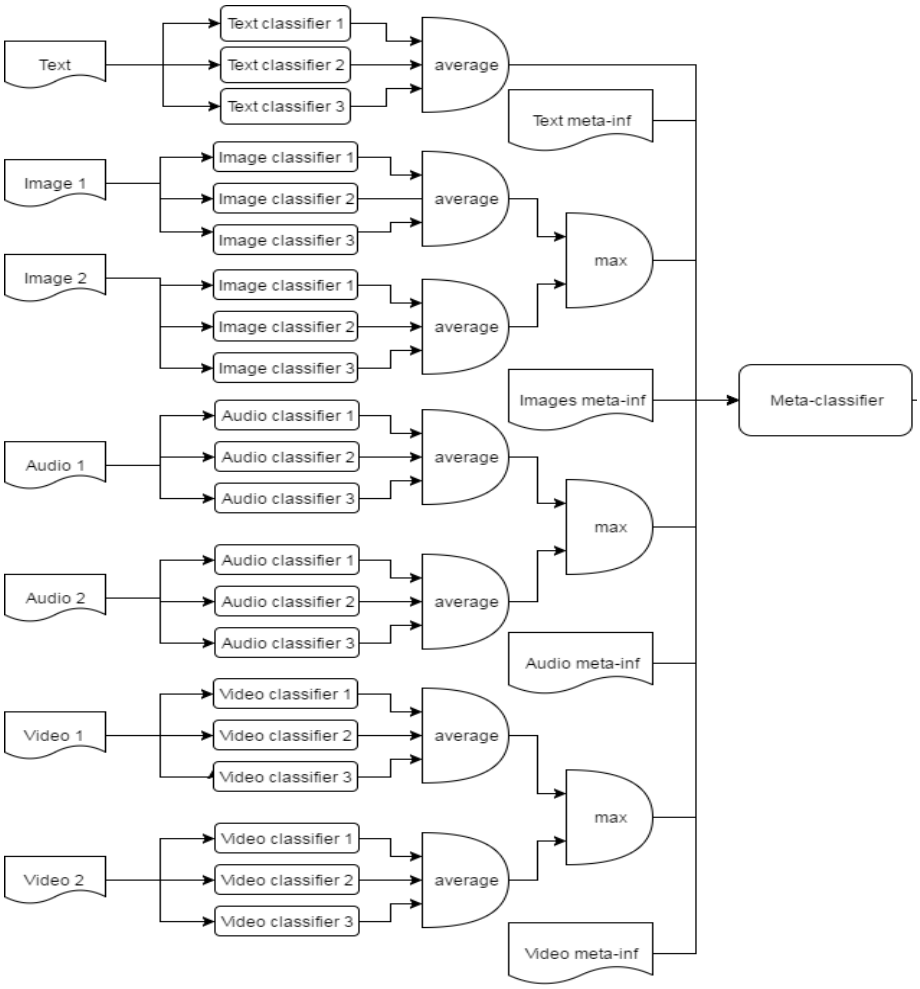

Рисунок 1. Комплексный классификатор информации

Для ранжирования информационных ресурсов предлагается использовать ансамбль классификаторов типа “Stacking” [16]. При использовании такого ансамбля строится иерархия классификаторов, в которой выходные значения классификаторов уровня ниже являются входными

данными классификаторов уровня выше. Структура ансамбля предполагается универсальной для всех материалов. Построение классификаторов для отдельных информационных ресурсов в работе рассматривать не будем.

Классификация каждого элемента (для изображений, аудио и видео элементов может быть множество) осуществляется произвольным подходящим набором классификаторов (единственное условие – для него должна существовать агрегирующая / сверточная функция). Как результат классификации элемента берется агрегированное значение оценок классификаторов (например, взвешенное среднее для алгоритмов бинарной классификации, выдающих байесовскую вероятность, или мода для алгоритмов со строгой оценкой).

Далее из этих оценок (за исключением оценки класса текста, так как задача сегментации текста и последующей его разметки предполагается слишком объемной [17], и при необходимости, должна быть рассмотрена отдельно) берется наибольшая, так как вполне вероятна ситуация, когда значительная часть элементов не содержат запрещенной информации. Максимальные агрегированные оценки каждого вида элемента подаются на вход мета-классификатора вместе с другой, заранее определенной, метаинформацией.

Под метаинформацией в комплексном классификаторе подразумевается любая информация, полученная как результат анализа информации. Таковой может являться любая информация, полученная в задачах выше. Однако следует учитывать, что задачи, обозначенные выше, преследуют цель дать пользователю дополнительную информацию для принятия решения. Кроме этой информации, комплексный классификатор может принимать на вход информацию, которая не может быть независимо интерпретируема пользователем.

Пример возможной структуры метаинформации приведен на рисунке 2:

Рисунок 2 Пример возможной метаинформации

Например, метаинформацией может выступать описание числа аудиофайлов на веб – странице: в случае, если аудиофайлов нет, то классификатор не будет учитывать оценку аудиофайлов; также поведение классификаторов будет различаться для разного числа аудиофайлов: например, небольшое количество аудиофайлов можно встретить на страницах пропаганды ислама (в дополнение к относительно объемному тексту), в то время как большое количество аудиофайлов может указывать на аудиохостинг либо веб – страницу с подборкой аудиозаписей, на которых, как правило, относительно малое количество релевантного текста. Как метаинформацию в вышеуказанном случае можно использовать например, единичный вектор, значения скаляров которого отображают попадание числа аудиозаписей в заранее определенный карман (например, карманы [0;0], [1;1], [2;7], [8;>8], которые заранее были эмпирически определены на основе статистических данных об оценках пользователей и количестве аудиофайлов). Рекомендуется выделить значимую с точки зрения классификатора метаинформацию и для других видов классифицируемых элементов.

Также, задача ранжирования выдачи предположительно может быть улучшена персонализацией выдачи на основе ранее обработанных экспертом материалов.

4. Модель сбора и выдачи информации

В модели сбора и выдачи информации определены следующие процессы:

1. Поиск материалов
2. Актуализация базы поисковых сущностей
3. Получение материалов по ссылке
4. Оценка материалов экспертом
5. Составление отчетов и статистики
6. Ранжирование выдачи эксперту

Процесс поиска материалов в сети «Интернет» предлагается реализовать с помощью поиска в поисковых системах (например, Google, Yandex, DuckDuckGo и т.д.) по запросам, сгенерированным на основе базы поисковых сущностей.

Процесс актуализации базы поисковых сущностей можно разделить на два процесса: актуализация базы экспертом через графический интерфейс пользователя, а так же автоматическая генерация запросов на основе анализа полученных ранее материалов. Для генерации новых поисковых запросов на основе анализа данных предлагается использовать те материалы, которые были оценены как содержащие противоправную информацию. Генерацию поисковых запросов возможно производить путем комбинации наиболее значимых слов в материалах [18]. Значимость слова можно определять, например, производя анализ распределений слова.

Получение материалов по ссылке предполагает получение содержимого веб-страницы в том виде, в котором оно бы наиболее вероятно отобразилось пользователю. Таким образом, лучше получать содержимое веб-страниц через браузер, например, Selenium.

Оценка материалов экспертом и составление отчетов и статистики в рамках доклада рассматривать нет необходимости, так как за исключением деталей технической реализации, сам процесс достаточно тривиален.

В процессе ранжирования ресурсов предлагается использовать результаты работы комплексного классификатора и извлеченных предварительных данных. В качестве рейтинга, например, возможно использовать следующее значение:

$$rating = ce + fd + fp + fr + fs$$

, где:

- ce – оценка классификатора, в интервале [0..1], где большее значение соответствует большей уверенности классификатора в наличии противоправной информации.
- fd – отношение дубликатов материала, оцененных экспертом как содержащие противоправную информацию, ко всем дубликатам материала. В случае если у материала нет дубликатов, полагается равным 0.
- fp – среднее отношение материалов, содержащих ту же самую личность из базы личностей, оцененных экспертом как содержащие противоправную информацию, ко всем материалам по этой личности, для всех личностей. В случае если нет выявленных личностей, полагается равным 0.
- fr – отношение материалов, содержащих противоправную информацию и размещенных на информационном ресурсе, ко всем материалам, размещенным на информационном ресурсе. В случае если нет оцененных материалов, полученных из данного информационного ресурса, полагается равным 0.
- fs – оценка содержания противоправной информации на основе извлечения дополнительных фактов (например, наличие определенной цепочки фактов, с большой долей вероятности определяющей наличие противоправной информации). Находится в интервале [0..1]. Механизм работы таких оценок будет сильно зависеть от контекста и вида информации. Описание работы такого механизма весьма объемно и выходит за рамки данной работы.
- Материалы для выдачи эксперту ранжируются по убыванию значения rating.

4. Заключение

В работе были рассмотрены следующие моменты:

- Определены проблемы поиска противоправной информации в сети «Интернет».
- Предложены возможные пути решения.
- Приведена модель сбора и выдачи информации.

Список используемых источников

1. Смирнов А. А. Роль и место средств массовой информации в механизме детерминации противоправного поведения //Библиотека криминалиста. Научный журнал. – 2012. – №. 1. – С. 288-299.
2. Кублин И. М., Тинякова В. И. Инструменты управления лояльностью пользователей в социальном медиа-маркетинге, их разновидности и функции //Поволжский торгово-экономический журнал. – 2013. – №. 5. – С. 56-62.
3. Волошин С. В. и др. Анализ качества бинарной классификации веб-страниц методом опорных векторов //Известия Алтайского государственного университета. – 2017. – №. 4 (96).
4. Карташев Е. А., Царегородцев А. Л. Автоматизированная информационная система поиска и анализа информации в сети Интернет //Фундаментальные исследования. – 2016. – Т. 2. – №. 10.
5. Kotelnikov E., Razova E., Fishcheva I. A Close Look at Russian Morphological Parsers: Which One Is the Best? //Conference on Artificial Intelligence and Natural Language. – Springer, Cham, 2017. – С. 131-142.
6. Сулейманов Р. С. Извлечение метаданных из полнотекстовых электронных русскоязычных изданий при помощи Томита-парсера //Программные продукты и системы. – 2016. – №. 4 (116).
7. Anh L. T., Arkhipov M. Y., Burtsev M. S. Application of a Hybrid Bi-LSTM-CRF model to the task of Russian Named Entity Recognition //arXiv preprint arXiv:1709.09686. – 2017.
8. Dernoncourt F., Lee J. Y., Szolovits P. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks //arXiv preprint arXiv:1705.05487. – 2017.
9. Soloshenko A. N. et al. Establishing Semantic Similarity of the Cluster Documents and Extracting Key Entities in the Problem of the Semantic Analysis of News Texts //Modern Applied Science. – 2015. – Т. 9. – №. 5. – С. 246.
10. Зеленков И. В., Сегалович И. В. Сравнительный анализ методов определения нечетких дубликатов для Web-документов //Труды 9ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». – RCDL’2007, Переславль-Залесский, Россия, 2007. – С. 166-174.

11. Joachims T. Text categorization with support vector machines: Learning with many relevant features //European conference on machine learning. – Springer, Berlin, Heidelberg, 1998. – С. 137-142.
12. Кафтанников И. Л., Парасич А. В. Особенности применения деревьев решений в задачах классификации //Вестник Южно-Уральского государственного университета. Серия: Компьютерные технологии, управление, радиоэлектроника. – 2015. – Т. 15. – №. 3.
13. Tang J. et al. Understanding the limiting factors of topic modeling via posterior contraction analysis //International Conference on Machine Learning. – 2014. – С. 190-198.
14. Potapenko A., Popov A., Vorontsov K. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks //Conference on Artificial Intelligence and Natural Language. – Springer, Cham, 2017. – С. 167-180.
15. Shi B., Bai X., Yao C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition //IEEE transactions on pattern analysis and machine intelligence. – 2017. – Т. 39. – №. 11. – С. 2298-2304.
16. Wolpert D. H. Stacked generalization //Neural networks. – 1992. – Т. 5. – №. 2. – С. 241-259.
17. Pasternack J., Roth D. Extracting article text from the web with maximum subsequence segmentation //Proceedings of the 18th international conference on World wide web. – ACM, 2009. – С. 971-980.
18. Hong K., Nenkova A. Improving the estimation of word importance for news multi-document summarization //Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. – 2014. – С. 712-721.