Подход к фильтрации запрещенного контента в веб-пространстве

© Е.А. Сидорова[1,2] © И.С. Кононенко[1,2] ©Ю.А. Загорулько[1,2]

[1] Институт систем информатики имени А.П. Ершова СО РАН,
[2] Новосибирский государственный университет,
Новосибирск, Россия

lsidorova@iis.nsk.su irina_k@cn.ru zagor@iis.nsk.su

Аннотация. Введение законодательного регулирования содержания информационных ресурсов обострило проблему автоматического обнаружения и блокировки запрещенного контента. Предложен подход к решению данной проблемы, в котором тематический анализ веб-сайтов дополняется жанровым, что позволяет выявить осуществляемую посредством веб-сайта деятельность и, благодаря этому, более точно распознать и локализовать запрещенный контент. Решение о наличии запрещенного контента на странице сайта принимается не только на основе анализа ее содержимого, но и на основе результатов анализа тематики и жанра сайта в целом. Разработаны программные средства и русскоязычные ресурсы для обнаружения запрещенного контента, относящегося к теме «Наркомания и наркотики».

Ключевые слова: классификация веб-сайтов, фильтрация запрещенного контента, тематический анализ текста, жанровый анализ веб-сайтов.

An Approach to Filtering Prohibited Content on the Web

© E.A. Sidorova[1,2] © I.S. Kononenko[1] © Yu.A. Zagorulko[1,2]

[1] A.P. Ershov Institute of Informatics Systems,
[2] Novosibirsk State University,
Novosibirsk, Russia

lsidorova@iis.nsk.su irina_k@cn.ru zagor@iis.nsk.su

Abstract. The institution of legislative regulation of the content of information resources has aggravated the problem of automatic detection and blocking of prohibited content. We propose an approach to solving this problem. In this approach, a thematic analysis of websites is complemented by a genre one, which allows identification of the activity carried out through a website and, therefore, brings about a more accurate recognition and localization of the illicit content. The decision on the presence of prohibited content on a website page is made on the basis of both analysis of the page text content and results of thematic and genre analysis of the site as a whole. Software and Russian-language resources for the detection of prohibited content related to the topic "Drug addiction and drugs" have been developed.

Keywords: website classification, filtering prohibited content, thematic text analysis, website genre analysis.

1 Введение

Задача избирательного распространения информации, сформулированная Луном (Luhn) в 1958 г., получила наименование «фильтрация» в 1975 г. (Denning). Система фильтрации контролирует поток документов, отбирая в нем полезные документы в соответствии с некоторым критерием (информационная потребность пользователя). Более полно задача определена в [5]: процесс фильтрации предназначен для отбора или удаления информации из динамического потока данных.

Введение законодательного регулирования содержания информационных ресурсов обострило проблему обнаружения и блокировки запрещенного контента, к которому относится любое запрещенное государством для просмотра и ознакомления информационное наполнение ресурса или веб-сайта (текст, мультимедиа, графика). При существующей скорости прироста и обновления информации в полной мере контролировать ее содержание с помощью модераторов-людей практически невозможно.

Современные подходы к автоматической фильтрации запрещенного контента чаще всего основаны на использовании списков ссылок на сайты (URL-фильтрация) [13], распознавании ключевых

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

слов из списка запрещенных, а также на основе тематической классификации, например [6, 10]. Указанные методы не дают требуемого качества: в первом случае списки составляются вручную и не позволяют оценивать новые сайты, во-втором случае ключевые слова дают очень грубую оценку и либо ложно блокируют сайты с употреблением терминов в других смыслах, либо недостаточно полно покрывают способы выражения запрещенной информации. Что касается тематической классификации, то, помимо большой зависимости от обучающей выборки, она не позволяет определить цели, с которыми дается та или иная информация, что приводит к ложному срабатыванию фильтра, а для огромных массивов интернет-данных это недопустимо.

При рассмотрении различных методов фильтрации [3, 5], таких, как Boolean Information Filtering, Vector Space Model, Neural Networks и т. п., подчеркивается важность семантических проблем, т. е. проблем неоднозначности терминов (синонимия, полисемия, омонимия), затрудняющих сопоставление терминов в процессе содержательной фильтрации. Для преодоления семантических проблем, например, в [7], предложен метод, основанный на лингвистической онтологии, в качестве которой используется WordNet [2]. Основным недостатком такого подхода является трудоемкость построения лингвистической онтологии для заданного языка и предметной области.

В предлагаемом нами решении используется комплексный подход, при котором решение о запрещенности страницы принимается на основании не только ее тематики, но и прагматики, т. е. вида деятельности, осуществляемой посредством сайта в целом. Дополнение тематического анализа жанровым, а также использование лексических признаков, позволяющих явным образом задать семантику терминов, дает возможность более точно распознать и локализовать запрещенный контент.

2 Задача фильтрации контента

Фильтрация текстового контента традиционно рассматривается как разновидность информационного поиска. С другой стороны, фильтрацию можно рассматривать как особый случай классификации по двум категориям (релевантные и нерелевантные). В обзоре [4] сформулированы сходства и различия информационного поиска, фильтрации и бинарной категоризации. Фильтрация, в отличие от поиска, основана не на запросах, а на представлении индивидуальных или групповых интересов (профиль пользователя). Запрос – сиюминутный интерес, а профиль – долговременный (возможно меняющийся) интерес.

Базовое сходство всех направлений заключается в наличии следующих компонентов:

1. Представление веб-объекта (документа).
2. Представление информационного класса (информационной потребности, категории, профиля пользователя).
3. Сопоставление документа и класса с помощью алгоритмов, вычисляющих меру сходства.

Запрещенный контент − это любое содержательное наполнение веб-сайта, предоставление которого для просмотра и ознакомления запрещено государством. На территории РФ действует федеральный закон № 149-ФЗ «Об информации, информационных технологиях и о защите информации», в соответствии с которым устанавливаются основания для включения сайтов в список запрещённых. Список тематик блокируемых ресурсов открыт и включает, к примеру, такие типы запрещенного контента, как: контент, предназначенный только для взрослых, пропаганда против отдельного лица, группы или организации; материалы, связанные с наркотиками; контент, связанный с оружием, и др.

Для апробации предлагаемого подхода в качестве запрещенного рассматривался текстовый контент на русском языке, относящийся к теме «Наркомания и наркотики».

В силу высокой сложности задачи выявления запрещенного контента предложенное решение основано на совокупности различных методов анализа текстов и интернет-документов, включая методы машинного обучения и инженерный подход.

Машинное обучение не является полностью автоматическим, оно также требует экспертной деятельности по аннотированию обучающего множества текстов метками классов. Однако сформированные автоматически (хотя и на основе экспертной разметки) описания классов содержат много «шумящей» лексики, которая на этапе классификации текстов понижает точность работы алгоритма.

Инженерный подход предполагает создание описаний классов с участием эксперта, который, используя ускоряющие его деятельность программные модули нормализации текста и генерации частотных словарей, формирует ресурсы для классификатора. Несмотря на трудоемкость реализации, инженерный подход обеспечивает высокое качество классификации текстов за счет экспертной фильтрации «шума» и дополнения словарей (описаний классов) недостающей лексикой, отсутствующей в обучающей коллекции.

Особенность предлагаемого решения состоит в интеграции тематических и жанровых методов классификации текстовых ресурсов на базе инженерных правил принятия решения о наличии вредоносного контента. Использование тематических градаций в теме «Наркомания и наркотики» обеспечивает построение ее описания во всем многообразии и полноту классификации контента.

Необходимость использования жанровой классификации вызвана особенностями основной

темы и требованиями к принимаемому решению – определению принадлежности контента к двум классам: запрещенному контенту и незапрещенному. Определение жанра позволяет уточнить решение, полученное на базе тематической классификации. Этому же способствуют используемые логические правила принятия решения о запрещенности контента, построенные на основе результатов жанровой и тематической классификации.

В силу особенностей текстов исследуемой тематики традиционный алгоритм обработки текстов дополнен модулем анализа специальной тематической и стилистически окрашенной лексики – научная терминология, сленг наркоманов, обесценная лексика, жаргон интернет-пользователей, тематическая лексика на латинице и транслите.

Для оптимизации времени работы приложения алгоритм реализуется в два этапа:

1. предварительный анализ: установление наличия в тексте лексики, характерной для заданной тематики;
2. основной алгоритм: тематическая и жанровая классификации с принятием окончательного решения о запрещенности / незапрещенности контента.

Предусмотрена возможность обоснования полученных решений путем предоставления промежуточных результатов работы алгоритма фильтрации в понятной для конечного пользователя форме: найденной лексики, полученной уточненной тематики, жанра и используемых решающих правил.

3 Модель знаний

Предлагаемое нами решение основано на использовании лингвистических и предметных знаний и включает следующие ресурсы:

1. Рубрикаторы: тематический, жанровый (жанры интернет-текстов), прагматический (жанры сайтов) и лексический (признаки терминов).
2. Предметный словарь, включающий тематическую и жанровую лексику.
3. Жанровые шаблоны веб-текстов.
4. Прагматические модели веб-сайтов.
5. Решающие правила.

Рассмотрим их подробнее.

Тематический рубрикатор вводит уточняющие подтемы для базовой тематики «Наркомания и наркотики» и включает как запрещенные темы, так и незапрещенные (см. Рис. 1).

Назначение данного рубрикатора:

- отделить сайты по заданной тематике;
- дать объяснение пользователю, почему сайт заподозрен или отнесен к запрещенным.

Жанровый рубрикатор предназначен для классификации веб-страниц и веб-сайтов по жанрам, что в дальнейшем используется как для уточнения тематической классификации, так и для повышения качества фильтрации на основе правил.

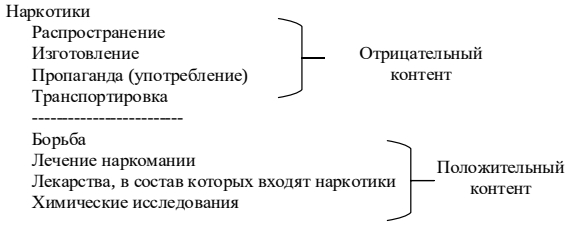

Рисунок 1 Фрагмент тематического рубрикатора

Выделяются такие жанры веб-ресурсов, как *Торговая площадка, Аптека, Сайт медицинской организации, Энциклопедический ресурс, Новостная лента, Персональная страница, Комментарий* и т. п.

Предметный словарь − структурированное хранилище терминов (слов и словокомплексов), в котором содержится вся необходимая информация для предварительного отбора тематически релевантных страниц, тематического и жанрового анализа текстового контента и принятия решения о блокировке.

Начальное наполнение словаря генерируется на этапе обучения с использованием размеченного экспертами корпуса веб-страниц, относящихся к исследуемой тематике, с применением универсального морфоанализатора, снабженного функцией предсказания незнакомой лексики.

Дополнительными источниками тематической лексики являются законодательно утверждённые Правительством РФ перечни наименований контролируемых наркотических средств, психотропных веществ и их прекурсоров, а также соответствующих видов растений, которые периодически пополняются и корректируются (примерно раз в год). Соответствующие документы доступны на официальных интернет-сайтах правовой информации, таких, как www.consultant.ru и pravo.gov.ru.

Далее осуществляется настройка предметного словаря экспертами, которые выделяют в его составе специальные подсловари, используя систему лексических признаков: тематическая лексика, научные термины, сленг наркоманов, термины на латинице, жанровая лексика и др. В задачу экспертов входит пополнение этих подсловарей, выявление регулярных ошибок фильтрации и формирование правил для изменения состава и структуры словаря.

Для создания и настройки словаря использовалась технология создания терминологических словарей KLAN [12].

Жанровые шаблоны веб-текстов формируются на основе лексических маркеров жанра и условий их встречаемости в текстовом фрагменте. Маркеры строятся на основе терминов словаря, при этом используются возможности представления совместной встречаемости терминов, альтернативности терминов в конкретной позиции (квазисинонимия), а также иерархической вложенности маркеров друг в друга. Например, страницы сайта типа *Торговая площадка* содержат следующие элементы:

- количественные конструкции (маркер: единица измерения “гр”, “мгр”),
- списки количественных конструкций (прайсы) с маркерами из жанровой лексики:

  Цены: 5гр. – 5 000 р, 10гр.— 9 000 р
- жанровая лексика: *цена, товар, закладка*.

Шаблон веб-страницы составляется из маркеров, на которые накладываются позиционные условия на тип фрагмента (заголовок, ссылка, выделенный текст, текст). Как и при описании маркеров, поддерживаются альтернативы и совместная встречаемость маркеров.

Рассмотрим для примера новостной шаблон:

«новостная лента»: [<_навигацияНовость, all_h>]

_навигацияНовость: [«главное за сутки»]

[«главное за сегодня»][«главное за день»]

[«все новости»][«основные новости»]

[«последние новости»][«лента новостей»]

Содержательно данный шаблон описывает следующее правило: если в одном из заголовков встретится один из маркеров группы *_навигацияНовость*, то это новостная лента.

Модель веб-сайта задается набором жанров веб-страниц, которые обязательно должны присутствовать на сайте и являются в совокупности его отличительным признаком. Для каждого сайта может быть задано несколько шаблонов. Например, модель интернет-магазина представлена двумя альтернативами:

[Магазин, Описание товара, ПредложениеТовара, Корзина, Доставка, Оплата]

[Магазин, Описание товара, ПредложениеТовара, СтатусЗаказа]

Принятие решения осуществляется на основе *решающих правил*, в посылках которых описываются условия того, будет ли анализируемый контент запрещен или разрешен. Эти условия строятся на термах, значениями которых являются конкретные тематики, жанры текста, жанры сайта и лексические признаки. Применяются правила двух видов: *положительные* и *отрицательные*, характеризующие текст, соответственно, как разрешенный или запрещенный. Правилами описываются, например, следующие экспертные наблюдения:

a) Если анализируемому контенту приписан лексический признак <40> «Обсценная лексика», он отнесен к тематике [601] «Употребление наркоманами» и жанру <401> «Торговая площадка» или (404) «Научная/информационная статья», то текст следует отнести к запрещенному контенту;

b) Текст по теме [1102] «Выращивание наркотических растений», написанный в жанре (407) «Словарная статья», относится к незапрещенному контенту. А текст по той же теме, представленный в ином жанре, может диагностироваться правилами как запрещенный контент и т. п.

Экспертные правила, помимо полноты, обладают высокой объяснительной способностью, что является существенным для нашей задачи.

Отметим, что правила принятия решений можно было бы сформировать автоматически при достаточном объеме обучающей выборки. Эксперимент показал, что экспертные правила не противоречат правилам, сформированным автоматически по обучающей выборке. Таким образом, можно рассматривать такой метод автоматического формирования правил как способ верификации правил, написанных экспертом.

4 Фильтрация контента

Анализ текстового контента осуществляется в несколько этапов. К основным этапам относятся тематическая и жанровая классификация текста, жанровый анализ сайта и принятие решения о запрещенности контента.

Объем статьи не позволяет в полной мере раскрыть каждый этап обработки текста, поэтому мы сконцентрируемся на основных идеях и используемых подходах.

4.1 Классификация текста

Прежде всего, необходимо уметь выявлять соответствие контента исследуемой тематике (*подозрительность* текста). При принятии решения о степени подозрительности контента необходимы:

а) Словарь тематической лексики, присутствие которой в тексте позволяет предположить тему «Наркомания и наркотики». Словарь содержит слова и словосочетания данного лексико-семанти-ческого поля, как специальные научные и нейтральные, так и жаргонные (сленг наркоманов). Эта лексика включает названия наркотиков, наркосодержащих лекарств и растений, названия состояний под воздействием наркотиков и т. п.

б) Критерий для определения возможной принадлежности к данной теме (степени подозрительности) текста, содержащего термины из словаря. Вычисление критерия опирается на степень присутствия тематической лексики с учетом лексического признака однозначности/ неоднозначности (омонимичная, т. е. тематически неоднозначная лексика из рассмотрения на данном шаге исключается).

Для подозрительных текстов применяется уточняющая классификация в соответствии с заданными рубриками с использованием весовых характеристик терминов, вычисляемых как ожидаемая взаимная информация (EMI) [9]. Данная мера позволяет оценить, сколько информации о классе – в теоретико-информационном смысле – содержит термин. Обучение и настройка алгоритма классификации производилась с участием эксперта.

При оценке релевантности текста классу (тематике) помимо веса термина учитывалась «зона текста», в которой встретился термин [1]: так, например, вес терминов в заголовках удваивался.

Способ взвешивания терминов, основанный на расчете EMI, дает улучшение на 5% по сравнению со способом взвешивания типа TF*IDF.

4.2 Жанровый анализ

В отличие от основной массы подходов к фильтрации, которые реализуют только контент-анализ страниц ресурсов, т. е. тематический анализ по ключевым словам, либо ограниченный жанровый анализ (преимущественно по формальным признакам, таким, как длина текста, количество букв, цифр и специальных признаков, количество ссылок и т. п. [8]), предложенный нами подход осуществляет многоаспектный жанрово-тематический анализ и классификацию. Используемые в рамках данного подхода признаки классификации явным или опосредованным образом отражают не только тематику анализируемых ресурсов, но и такие коммуникативно-прагматические аспекты жанра, как вид деятельности, осуществляемой посредством ресурса, включая цели и задачи деятельности и целевую аудиторию как ее участника, медийные свойства ресурсов, стилистические особенности используемых языковых средств.

Признаки жанрово-тематической классификации делятся на группы, каждая из которых отражает определенный аспект классификации:

1. Жанрово-структурная классификация ресурсов на основе двухуровневой модели:
   - Макроуровень – ресурс в целом;
   - Микроуровень (компоненты ресурса: страница, раздел, блок).
4. Жанрово-прагматическая классификация ресурсов (на основе прагматических аспектов содержания и представления):
   - Праксиологические (деятельностные) аспекты (вид деятельности, которая осуществляется посредством ресурса);
   - Аспекты содержания и представления, связанные с каналом коммуникации (медийные свойства ресурсов).
5. Жанрово-стилистическая классификация ресурсов:
   - Лексико-стилистические аспекты содержания и представления (стилистические особенности используемых языковых средств с акцентом на стилистически окрашенные языковые средства).

Представление о жанре закладывается на этапе формирования обучающей выборки, которая целенаправленно отбирается и размечается экспертами. Предлагаемая процедура жанровой классификации совмещает статистический и экспертный подходы к анализу жанра и опирается на метод вычисления меры принадлежности текста к жанру [11]. Вначале применяется экспертный подход, в рамках которого осуществляется поиск в тексте жанровых маркеров, т. е. сопоставление тексту шаблонов, составленных экспертом. Если на основе маркеров жанр веб-текста определить не удалось, то применяется классификация на основе методов машинного обучения.

4.3 Принятие решения на основе правил

Решение о запрещенности/незапрещенности контента принимается на основе следующих параметров:

1. $\bar{P}_t$=($p(t_1)$, $p(t_2)$, ..., $p(t_i)$, ..., $p(t_{Nt})$) – вектора релевантности текстового контента тематикам рубрикатора, где Nt – число тематик в рубрикаторе, $p(t_i)$ – вероятность реализации тематики t_i в анализируемом тексте, i = 1, ..., Nt; $\sum_{i=1}^{Nt} p(t_i) = 1$;
2. $\bar{P}_j$ =($p(j_1)$, ..., $p(j_{Nj})$) – вектора релевантности контента текста жанрам текста, заданным в жанровом рубрикаторе, где Nj – число жанров текста в рубрикаторе; $\sum_{i=1}^{Nj} p(j_i) = 1$;
3. $\bar{P}_{js}$ =($p(js_1)$, ... ,$p(js_{Ns})$) – вектора релевантности контента всего сайта жанрам сайта, заданным в рубрикаторе, где Ns – число жанров сайта в рубрикаторе; $\sum_{i=1}^{Ns} p(js_i) = 1$;
4. $VL=(v(lex_1), \ldots, v(lex_{Ln}))$ – вектора наличия лексических признаков в текстовом контенте, где $v(lex_i) \in \{0,1\}$ – показатель присутствия/ отсутствия в тексте лексического признака lex_i (например, сленга, обсценной лексики и т. п.);
5. $\bar{P}_{Rule}$ – набора решающих правил вида t_i & j_k & js_m & lex_j, принимающих решение о запрещенности / незапрещенности анализируемого контента в виде оценки m^p, вычисляемой как вероятность совместной реализации темы t_i, жанра текста j_k, жанра сайта js_m и лексического признака lex_j. в этом контенте. Оценка m^p вычисляется по формуле $p(t_i) \cdot p(j_k) \cdot p(js_m) \cdot v(lex_j)$, т. е. это произведение вероятностей указанных в правиле параметров, взятых из векторов, описанных выше;
6. $\bar{M} = (M^-, M^+)$ – двухкомпонентный вектор сумм оценок всех отрицательных и положительных правил соответственно.

Окончательное решение о запрещенности / незапрещенности контента принимается по критерию C: если $C = (M^- - M^+) > 0$, то считается, что контент запрещен. Настройка данного критерия позволяет изменять результаты работы системы в сторону повышения либо полноты, либо точности фильтрации.

5 Архитектура системы фильтрации запрещенного контента

Схема выявления запрещенного контента представлена на Рис. 2. На вход системы фильтрации запрещенного контента поступает контент сайта, представленный множеством веб-текстов (текстов с html-разметкой), либо обновление сайта – множество новых либо отредактированных веб-текстов сайта. Веб-текст – это единица текстового контента сайта, хранящаяся в БД на сервере. Веб-страница, которую видит пользователь при просмотре веб-сайта с

помощью веб-браузера на стороне клиента, формируется в общем случае из множества веб-текстов с добавлением незначащего для анализа контента – элементов оформления страницы, баннеров, рекламы и т. п., а также медиа-контента.

Обработка сайта начинается с анализа его структуры, затем формируется начальный индекс сайта (в случае обновления сайта индекс модифицируется), фиксируются зависимости между веб-текстами. После этого тексты сайта последовательно анализируются.

Рисунок 2 Схема выявления запрещенного контента

Каждый веб-текст очищается от html-разметки (значащие элементы разметки, такие, как заголовки, ссылки, выделение фрагмента стилем, сохраняются), осуществляется лингвистический анализ текста, обеспечивающий поиск в нем терминов словаря, и сбор статистической информации. Далее производится оценка тематической принадлежности текста к базовой теме «Наркомания и наркотики» – т. н. «оценка подозрительности» текста (текст считается подозрительным, если его контент соответствует базовой теме). В определении подозрительности участвует только однозначная лексика, наличие которой позволяет снять возможную тематическую неоднозначность текста. Для неподозрительных текстов дальнейшая оценка запрещенности не проводится, определяется лишь жанр текста, который заносится в индекс сайта.

Жанровая классификация позволяет определить жанр текста на основе словаря маркеров и структурного анализа текста в соответствии с разметкой. Если на основе маркеров и жанровых шаблонов жанр веб-текста определить не удалось, то применяется уточняющая классификация на основе методов машинного обучения.

Уточняющая классификация обеспечивает не только определение жанра текста, но и уточнение (конкретизацию) его тематики в соответствии с типами противоправных и разрешенных действий в рамках темы «Наркомания и наркотики». При уточняющей классификации используется обученный на размеченном корпусе текстов предметный словарь. Результатом уточняющей классификации являются векторы релевантности текста темам и жанрам, которые сохраняются в индексе сайта.

После первичной обработки всех веб-текстов сайта осуществляется анализ его жанра. Каждый жанр сайта описывается одной или несколькими моделями. Модель сайта фиксирует набор жанров текста, которые обязательно должны встретиться на сайте данного жанра. Данные модели составляются экспертами вручную на основе анализа структуры веб-сайтов обучающей коллекции. Вычисление оценки степени соответствия сайта какому-либо жанру осуществляется по моделям сайтов и оценкам, полученным для жанров веб-текстов сайта. Полученные оценки для жанра веб-сайта и составляющих его веб-текстов сохраняются в индексе сайта.

Принятие решения о запрещенности сайта осуществляется на основе решающих правил, которые применяются только для подозрительных текстов. Особенностью параметра подозрительности текста является то, что он «распространяется» на все связанные тексты (связи между текстами фиксируются структурой сайта и хранятся в индексе сайта). Поэтому на стадии предварительной обработки осуществляется поиск всех подозрительных текстов по связям и выполнение уточняющей классификации для тех из них, для которых она ранее не проводилась. Результатом применения правил к тексту является оценка запрещенности страницы.

Оценка запрещенности всего сайта определяется как максимум из оценок запрещенности по всем текстам сайта.

6 Результаты эксперимента

Для оценки качества фильтрации были сформированы одна обучающая и две тестовых коллекции, содержащие веб-тексты:

1. Обучающая коллекция, состоящая из 468 веб-текстов на русском языке, относящихся к теме «Наркомания и наркотики». Все тексты размечены экспертами. Разметка включает экспертную оценку запрещенности / незапрещенности контента, тематику, жанр веб-текста и жанр веб-сайта, на котором был размещен данный текст.
2. Тестовая коллекция веб-текстов, включающая около 123 тыс. русскоязычных веб-страниц, часть которых относится к теме «Наркомания и наркотики», но не содержит запрещенный контент.
3. Коллекция собрана вручную на основе сайтов Яндекс-каталога (https://yandex.ru/yaca). Тестовая коллекция веб-текстов, включающая 569 веб-текстов на русском языке, содержащих

запрещенный контент по теме «Наркомания и наркотики».

Полученные коллекции включают веб-тексты различных функциональных стилей – от нормативных и официальных документов до сообщений и комментариев на форумах и в социальных сетях, – что позволяет адекватно оценить качество фильтрации на всем многообразии интернет-жанров. К сожалению, в открытом доступе отсутствуют размеченные коллекции текстов по данной тематике, чем объясняется небольшой объем первой и третьей коллекций, которые создавались нашими экспертами вручную. Объем веб-текстов в коллекциях варьировался от 213 до 65655 Кб.

На основе обучающего корпуса текстов был построен словарь, который в дальнейшем был дополнен терминами из специализированных словарей. Словарь содержит более 50 тыс. терминов (без учета стоп-слов). Его общий количественный и качественный состав отражен в Таблице 1.

Таблица 1 Терминологический состав словаря

Лексем	Слово-комплексов	Подозри-тельных	Жанро-вых	Сленг
24175	26540	5349	1895	3161

Как видно из таблицы 1, ключевые слова для предварительного отбора текстов по теме («подозрительные», т. е. однозначные тематические термины) составляют десятую часть объема словаря.

Оценка качества классификации была дана в виде показателей полноты (R), точности (P) и F-меры. Рассматривалась бинарная классификация (1) и уточняющая тематическая классификация (2). Оба сравниваемых метода основаны на машинном обучении, но во втором случае используется расширенный набор тем, причем для каждой из них указано, является ли она запрещенной или нет.

Таблица 2 Сравнение методов классификации

	R	P	F-мера	Скорость
(1)	52,0%	65,4%	57,9%	~ 0,07 мс
(2)	72,6%	69,7%	71,1%	~ 0,10 мс

Как видно из Таблицы 2, использование уточняющего тематического рубрикатора, построенного по специальной ориентированной на задачу фильтрации методике, позволило улучшить показатели полноты и точности в сравнении с бинарной классификацией (когда контент сразу классифицируется на два класса – запрещенный и незапрещенный), соответственно, на 20% и 10%. Однако эти показатели все еще являются низкими.

Таблица 3 Оценка качества фильтрации

	Кол-во (страниц)	Правильных ответов (%)
Нейтральная коллекция	~ 123 тыс.	99.4%
Отрицательная коллекция	569	86.99%

В Таблице 3 приведены оценки работы системы фильтрации, в которой тематическая классификация сочетается с жанровой и применяются решающие правила (отметим, что результаты, полученные тематическим классификатором, использовались здесь в качестве промежуточных.)

Таким образом, ошибка первого рода составила 0,6%, ошибка второго рода – 13,01%.

Большая часть ошибок обоих типов связана с неполнотой словаря. Так, возможны существенные лакуны в подсловарях латиницы и транслита (например, отсутствуют названия наркотиков *25i-nbome, JWH, нбоме, дживиаш*). Не всегда в словаре учтена возможная лексическая или лексико-морфологическая неоднозначность (например, *доб.* может представлять в тексте наркотик или сокращение от *добавочный*).

Ложно-положительная оценка характерна для страниц, которые не проходят предварительный этап фильтрации ввиду отсутствия однозначной тематической лексики. Так, не блокируются (отсеиваются как неподозрительные) страницы, содержащие предложения или рекламу наркотических веществ, завуалированные путем использования неоднозначной лексики (например, *соли для ванн*), а также намеренно искаженные (зашифрованные) тексты.

Ложно-отрицательная оценка характерна для следующих типов веб-текстов: а) информационные статьи о наркотических веществах или растениях (в частности, о выращивании декоративных растений), жанр которых не определен как энциклопедическая/ словарная статья; б) новостные тематические тексты с позитивной окраской (*Умеренное потребление алкоголя и амфетамина может улучшить память у пожилых людей*); в) тематически нейтральные страницы комментариев на форумах и в блогах с вкраплением шутливых тематических комментариев (*Наркотой там не барыжите, случайно?* – реплика при обсуждении вопросов информационной безопасности).

Заключение

Предложенный подход реализован в виде приложения, интегрированного в платформу Plesk. Приложение позволяет выявлять и блокировать сайты, содержащие запрещенную информацию по теме «Наркомания и наркотики» и/или осуществляющие незаконную деятельность по торговле, распространению, транспортировке, изготовлению и пропаганде наркотиков.

К преимуществам предложенного подхода относятся, во-первых, глубокий анализ текстового контента веб-ресурса с учетом его тематических и жанровых особенностей, во-вторых, совмещение статистических и инженерных методов анализа текста, в частности, предложен уникальный метод принятия решения о запрещённости контента на основе решающих правил, учитывающих результаты его жанровой и тематической классификации, в-третьих, масштабируемость и технологичность разработанных программных средств, что позволяет

легко адаптироваться к различным предметным областям посредством настройки базы знаний.

В предложенном подходе, на наш взгляд, достигнут баланс между ручной работой эксперта и автоматическим обучением, где, во-первых, словари создаются и обучаются автоматически, а эксперты пополняют их номенклатурными терминами и сленгом, во-вторых, неполнота жанровых (функциональных) описаний интернет-ресурсов (создаются экспертом) компенсируется поддержкой статистического жанрового классификатора, и наконец, решающие правила потенциально могут строиться автоматически, а оценка применимости правила для каждого конкретного случая оценивается по вероятностной формуле.

Дальнейшее развитие описанной технологии связано с необходимостью автоматизации поддержки словаря в актуальном состоянии. Автоматизация возможна на базе жанрового анализа страниц, относящихся к жанрам «Нормативный список» (отслеживание словарей официальных наименований контролируемых веществ и растений) и «Словарная статья» (отслеживание словарей универсального и тематического сленга, обсценной лексики). Однако главным источником тематической лексики по-прежнему остаются эксперты, т. к. интернет-словари тематического сленга существенно отстают от происходящих в среде наркоманов изменений лексики.

В качестве актуального направления исследований по данной тематике также рассматривается возможность применения методов сентимент-анализа для улучшения распознавания трудноуловимой темы пропаганды наркотиков, представленной в информационных сообщениях, создающих привлекательный образ наркомана и процесса употребления наркотических веществ.

Благодарности

Работа выполнена при финансовой поддержке Министерства образования и науки Российской Федерации (договор № 02.G25.31.0054) и Российского фонда фундаментальных исследований (грант № 15-07-04144).

Литература

[1] Cohen, William W., Singer, Y.: Context-sensitive Learning Methods for Text Categorization. ACM Transactions on Information Systems, 17, pp. 141-173 (1999)

[2] Gormez, Josre M., Girarldez, I., De Buenaga, M.: Text Categorization for Internet Content Filtering. Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial, 920, pp. 34-52 (2003)

[3] Khozooii, N.S., Haratizadeh, S., Keyvanpour, M.R.: An Analytical Framework for Web Information Filtering Techniques. Int. J. of Hybrid Information Technology, 6 (6), pp. 345-358 (2013)

[4] Nanas, N.: Literature Review: Information Filtering for Knowledge Management. The Open University, 2001. http://kmi.open.ac.uk/publications/ pdf/kmi-01-16.pdf

[5] Nouali O., Blache P. Automatic Classification and Filtering of Electronic Information: Knowledge-Based Filtering Approach. Int. Arab J. of Information Technology, 1 (1), pp. 85-92 (2004)

[6] Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Computing Surveys, 34 (1), pp. 1-47 (2002)

[7] Shoval, P., Maidel, V., Shapira, B.: An Ontology Content-based Filtering Method. Int. J. Information Theories & Applications, 15, pp. 303-314 (2008)

[8] Воронов, С.О., Воронцов, К.В.: Автоматическая фильтрация русскоязычного научного контента методами машинного обучения и тематического моделирования. Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Межд. конф. «Диалог». 2015. http://www.dialog-21.ru/digests/dialog2015/materials/pdf/VoronovSOVorontsovKV. pdf

[9] Маннинг, К.Д., Рагхаван, П., Шютце Х.: Введение в информационный поиск. М.: Вильяме, 528 с. (2011)

[10] Патент РФ № 2446460, МПК G06F21/20. Способ и система фильтрации веб-контента /Осипов Г.С., Тихомиров И.А., Соченков И.В.; патентообладатель ИСА РАН; заявл. 2010-11-18; опубл. 27.03.2012

[11] Сидорова, Е.А., Боровикова О.И.: Подход к жанровой классификации текстовых ресурсов. Информационные технологии и системы [Электронный ресурс]: Тр. Шестой Межд. науч. конф. ИТиС–2017: науч. электрон. изд. / отв. ред. Ю.С. Попков, А.В. Мельников. Челябинск: Челяб. гос. ун-т, сс. 264-269 (2017)

[12] Сидорова, Е.А.: Подход к построению предметных словарей по корпусу текстов. Труды межд. конф. «Корпусная лингвистика – 2008». СПб.: СПбУ, Факультет филол. и искусств, сс. 365-372 (2008)

[13] Стрекалов, И.Э., Новиков, А.А., Лопатин, Д.В.: Система формирования безопасности контента. Вестник ТГУ, 20 (2), сс. 462-464 (2015)