УДК 004.853
DOI: 10.20310/1810-0198-2017-22-5-1133-1137

СИСТЕМА ФОРМИРОВАНИЯ БЕЗОПАСНОГО КОНТЕНТА

© Д.В. Лопатин, Е.С. Чиркин
Тамбовский государственный университет им. Г.Р. Державина
392000, Российская Федерация, г. Тамбов, ул. Интернациональная, 33
E-mail: ccs.tmb@ya.ru

Приведено описание системы формирования безопасного контента для различных категорий пользователей. Для надежного детектирования негативного содержания выделены сущности корпусы нежелательного контента из большого массива текстов. Система формирования безопасного контента построена по клиент-серверной архитектуре и включает в себя клиентскую часть в форме расширения браузера. Серверный модуль имеет функцию загрузки пользовательского контента для последующего поиска на основе апробированных методов (наивный поиск, поиск регулярными выражениями, метод шинглов с рядом модификаций, фонетический поиск) или их комбинацией. Система модифицирует на стороне пользователя нежелательный контент путем полного или частичного блокирования (замены) содержание веб-страницы.
Ключевые слова: система; контент; безопасность; пользователь; родительский контроль

ВВЕДЕНИЕ

В настоящее время сеть Интернет характеризуется наличием большого количества разнообразной негативной информации (прямые и косвенные оскорбления по расовому, половому, социальному, групповому признаку, введение в заблуждение, шантаж и манипуляции, побуждения к действию или бездействию, распространение конфиденциальной информации, пропаганда аддиктивных зависимостей (алкоголь и наркотические вещества)), как при межличностном общении, так и направленной неограниченному кругу лиц, что влечет за собой необходимость ее своевременной идентификации и, возможно, блокирования. При этом размеры текстового контента на веб-странице незначительны. Однако многообразие способов выражения и форм представления контента на реальной веб-странице (словари эвфемизмов, субкультурного сленга, сокращений, ошибок в написании и другой нежелательной информации имеют значительные размеры) требует комплексного решения на основе алгоритмов поиска. Цель работы: разработать систему формирования безопасного контента для различных категорий пользователей.

МОДИФИКАЦИЯ КОНТЕНТА НА СТРАНИЦЕ

Определим две группы правил модификации контента на странице. К первой группе отнесем полное блокирование веб-страницы с отображением сообщения об ошибке и/или перенаправление пользователя на образовательный ресурс. Методы полного блокирования наилучшим образом подходят младшей возрастной группе пользователей; учреждениям, занимающимся обучением несовершеннолетних лиц; новым пользователям сети Интернет, которые не имеют навыков распознавания и блокирования угроз, связанных с потреблением негативного контента. Ко второй группе отнесем методы, которые предполагают частичное удаление контента или изменение графических свойств страницы (точечное удаление слов или абзацев с негативным содержанием, изменение свойств отображения негативного контента, частичная подмена содержимого веб-страницы обучающим контентом). Использование метода точечного удаления слов может быть полезно для пользователей, не желающих видеть случайно встречающейся негативный контент на страницах. Если невозможно нарушить контекстную связь внутри абзаца, можно удалить всю структурную единицу веб-страницы. Возможна замена удаляемого контента на обучающий блок, сформированный в зависимости от содержания удаленного фрагмента. Метод наилучшим образом подойдет для пользователей, которые согласны на модификацию нежелательного контента в целях обучения. Метод, основанный на изменении свойств отображения негативного контента, наилучшим образом подходит для пользователей, которые самостоятельно хотят повысить свой уровень знаний в области информационной безопасности и самостоятельно контролировать сетевые ресурсы.

КЛАССИФИКАЦИЯ СУЩНОСТЕЙ НЕЖЕЛАТЕЛЬНОГО КОНТЕНТА

Объем словаря нежелательных терминов является важной частью системы поиска, т. к. от него в первом приближении зависит точность и скорость работы всей системы детектирования нежелательного контента. Отсюда следует, что к вопросу формирования реальных словарей необходимо подходить с особой тщательностью. При этом задача усложняется по причине того, что условию задачи необходимо проводить анализ контента веб-страницы произвольной, неизвестной заранее, структуры. Исходя из предположения, что нежелательный контент или языковая единица достаточно четко определен, например, нецензурная лексика

в ее формальном определении, влекущем правовые последствия для российских средств массовой информации, то задачу его фильтрации можно свести к задаче обучения классификатора документов с целью выделения сущностей из текста, а также использовать это предположение при применении алгоритмов поиска. При этом сущности могут быть заранее известны (нежелательный контент), написаны с большим количеством ошибок, переданы в любой искаженной форме. Решение данной задачи дает возможность тонкой настройки практически любого алгоритма нечеткого поиска, поскольку выделение сущностей, сопутствующих искомым языковым единицам, восстанавливает контекст и позволяет выявить наличие последних даже при сильно искаженных формах написания. При этом механизм выделения сущностей нежелательного контента позволяет зафиксировать разумную степень глубины поиска и, таким образом, сэкономить вычислительные ресурсы, не осуществляя поиск наиболее редких вариантов написания. Размеченных корпусов нежелательного контента в свободном доступе не найдено. Для работы был создан и размечен корпус текстов из порядка 400000 слов, содержащих порядка 17000 упоминаний сущностей «популярного» нежелательного контента. Источником для разметки послужили, в основном, комментарии пользователей ряда популярных среди молодежи сетевых развлекательных ресурсов. Данные ресурсы слабомодерируемые и практически не имеют цензурных комментариев.

Классификация выполнялась в три этапа: на первом этапе производится классификация документов с использованием простейшего байесовского классификатора (относится текст к одному из классов по производящей основе слова или нет); на втором этапе идет уточнение, какой именно фрагмент документа относится к производящей основе слова; на третьем этапе идет оценка результата обучения, и в случае, если он хуже порогового значения, второй этап повторяется.

Размеченные документы разделялись однородно на обучающую и тестовую выборки в пропорции 70 %:30 %. Затем на обучающей выборке тренировалась модель до достижения количества ошибок по тестовой выборке менее 10 %, полученная модель применялась ко всему неразмеченному массиву документов. В результате применения были выделены сущности, сопутствующие искомым языковым единицам: в основном, это были служебные части речи и фрагменты обстоятельств места либо времени. Было замечено, что вероятность того, что при количестве комментариев к записи на любом ресурсе от 10 хотя бы один из комментариев будет с нецензурным словом или выражением, составляет величину, близкую к 100 %. Количество ошибок на тестовой выборке снижается с 17 % при размере обучающей выборки в 50 примеров до 0,01 % при выборке в 17000 примеров (табл. 1).

Величина ошибки для наивного поиска в зависимости от размера выборки (табл. 1) хорошо демонстрирует ситуацию, когда обучающая выборка перекрывает все возможные тестовые случаи. В любых нечетких алгоритмах это называется «переобучением» и является одной из ошибок обучения, однако данное понятие к точному поиску неприменимо и показывает лишь его превосходство над остальными алгоритмами в данных условиях, что стало возможным благодаря сочетанию двух факторов: большая обучающая выборка и ограниченное количество основополагающих корней, от которых произошли составляющие всех примеров, а также ряду ранее упомянутых ограничений. Алгоритм фонетического поиска демонстрирует некоторое превосходство над точными методами поиска при малых объемах обучающей выборки, это объясняется тем, что «приблизительность» алгоритма формирования фонетического алфавита покрывает недостаточность количества примеров, однако с ростом размера обучающей выборки это становится негативным фактором в связи с ростом количества ложноположительных случаев детектирования. Метод шинглов неприменим при малых объемах обучающей выборки. Ситуация исправляется при увеличении размеров выборки, поскольку в ней увеличивается доля сложносоставных слов, словосочетаний и иных случаев, для которых данный алгоритм является более приемлемым.

Таблица 1

Величина ошибки в тестовой выборке

Размер обучающей выборки, примеров	Величина ошибки, %		
	Наивный поиск	Фонетический поиск	Метод шинглов
50	20,79	10,11	89,75
100	10,16	3,72	74,32
500	2,12	2,64	61,84
1000	1,01	2,55	44,86
2000	0,52	2,61	31,52
3000	0,33	2,72	18,74
4000	0,21	2,61	13,92
5000	0,2	2,54	9,46
6000	0,16	2,03	8,53
7000	0,15	1,45	6,99
17000	0,01	0,11	1,01

СИСТЕМЫ ФОРМИРОВАНИЯ БЕЗОПАСНОГО КОНТЕНТА

Система формирования безопасного контента построена по клиент-серверной архитектуре и включает в себя клиентскую часть в форме расширения (chrome extension) для наиболее распространенного в настоящее время браузера на базе Chromium. Серверная часть состоит из нескольких слабосвязанных модулей на языке PHP. Выбор платформы клиентской части – расширения для браузера Chromium продиктован рядом причин: а) необходимо формирование безопасного контента независимо от площадки, его размещающей (взаимодействие с владельцами площадки исключено по очевидным причинам); б) эта система предназначена для конечного пользователя; в) это наиболее распространенный браузер (от ~65 до ~85 % от доли всех браузеров на рынке, в зависимости от методики подсчета и региона); г) значительная часть сетевых ресурсов использует или декларирует переход на зашифрованный протокол HTTPS (или HTTP/2.0) для обмена информацией с пользователем, что делает невозможных вмешательство в трафик. Использование расширения браузера в таком случае является единственным способом для мониторинга и модификации передаваемой информации.

Клиентская часть в виде расширения для браузеров на основе Chromium, осуществляющая фильтрацию

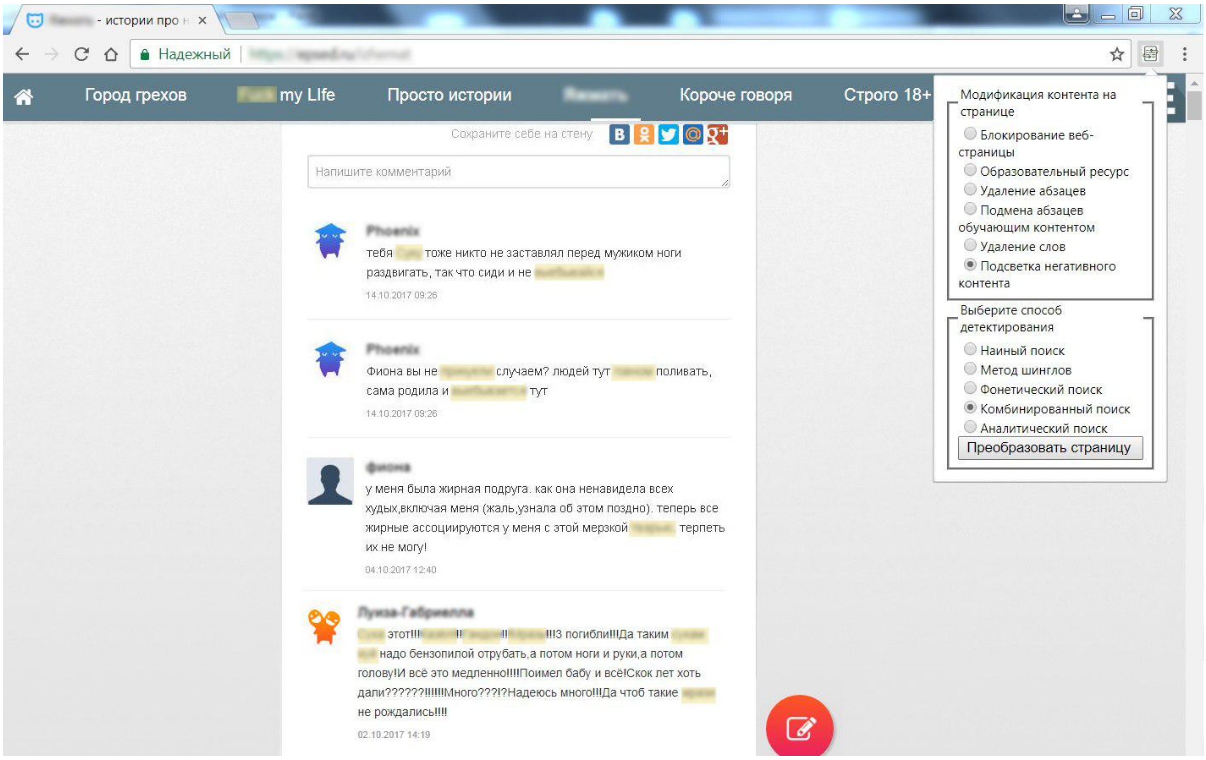

Рис. 1. Пример работы системы доставки безопасного контента

контента на странице: предварительная и/или ограниченная – посредством встроенных набора правил и алгоритмов правил модификации контента, и полнофункциональная фильтрация через обращение к серверному модулю и трансляция результата. На рис. 1 показан пример работы клиентской части системы доставки безопасного контента, пользователь выбрал акцентирование нежелательного контента (отметим, что данный вариант – наиболее популярный среди студентов).

Расширение построено по типовой схеме: popup-модуль – для взаимодействия с конечным пользователем (в данном случае – авторизации в системе) и информационных сообщений, предоставления краткой справки; content-модуль – внедряется в просматриваемую пользователем страницу, осуществляет ее первичную фильтрацию (на случай потери связи с серверной частью системы), а также устанавливает обработчики, следящие за обновлением контента на странице (согласно бизнес-логике сайта) и которые будут проверять обновившиеся части просматриваемого сайта; event-page-модуль предназначен для взаимодействия content-модуля и серверной части (они не могут взаимодействовать напрямую в большинстве возможных случаев). Помимо этого, eventpage-модуль является общим хранилищем ресурсов, необходимых для проведения фильтрации (настроек, стилей, функций, шаблонов, исключений), что позволяет значительно экономить вычислительные ресурсы при анализе, когда пользователь пассивно просматривает веб-страницу.

Серверная часть состоит из ряда слабосвязанных модулей на языке PHP. Модуль «API» осуществляет фильтрацию трафика согласно запросам клиентской части, ведет учет статистики. Модуль «сайт» носит обучающий и справочно-информативный характер; содержит инструкции по установке, настойке и использованию системы. Административный модуль предназначен для управления пользователями, настройки уровня фильтрации (белый и черный списки, с глубиной проверки), уровень замещения или пояснения, демонстрационный и тестовый разделы, раздел управления правилами, а также раздел анализа статистики. Административный модуль имеет возможности тонкой настройки уровня фильтрации: белый и черный списки сайтов, слов и словосочетаний, частные настройки каждого из реализованных алгоритмов.

Таблица 2

Максимальная задержка окончательного формирования страницы для конечного пользователя, вызванная использованием системы фильтрации контента на стендовом испытании

Размер тестовой выборки, примеров	Время ответа сервера, мс			
	Наивный поиск	Метод шинглов	Фонетический поиск	Комбинированный поиск
50	4	843	14	11
100	4	790	25	11
500	13	934	128	16
1000	67	865	142	160
2000	643	891	124	164
3000	570	754	132	162
4000	765	765	138	170
5000	832	857	136	166
6000	570	788	138	172
7000	840	894	142	160

Отзывчивость и быстродействие, а также незаметность системы для конечного пользователя можно весьма точно оценить по расходам системных ресурсов, которые потребляет приложение. В данном конкретном случае важным параметром можно назвать некую условную величину «время ответа сервера», которая приведена в табл. 2 для тестовой конфигурации. Следует отметить, что тестовая конфигурация имитирует худший из типовых случаев – соединение типа «медленный 3G» и сервер, не являющийся таковым по предназначению (разрешение DNS ~110…130 мс, сетевая задержка установления соединения ~540...650 мс, задержки запроса и ответа сервером – по ~40...90 мс, скорость соединения от ~100 кбит/с), в реальной эксплуатации все сетевые и серверные задержки будут на порядок ниже. Наивный поиск является основой предварительной фильтрации и сначала в уменьшенном варианте исполняется на клиенте (задержка ~4...100 мс (строки 1–4)) и полностью на сервере – время исполнения до 40 мс, время интерпретации на клиенте ~4 мс и меньше, остальное – сетевые задержки (строки, начиная с 5-й, табл. 2).

Хорошо видна граница в 2000 сэмплов, начиная с которой задействована серверная часть системы фильтрации (до – не используется, поскольку результаты одинаковы). Метод шинглов на клиенте не реализовывался ввиду его непроизводительности при малых объемах словаря, время исполнения составляет ~80–85 мс на сервере, весь остальной вклад вносят сетевые задержки. Фонетический анализ ввиду особенностей его реализации имеет время задержки – линейное по отношению к длине объема проверяемой веб-страницы, однако, при среднем размере страницы до 120 кбайт оно не оказывает заметного влияния по сравнению задержкой при подготовке внутренних структур к поиску. Фонетический анализ возможно целиком реализовать на клиенте и периодически обновлять с сервера необходимые словари. Исходя из этих предпосылок был разработан комбинированный алгоритм (с дополнительный задержкой ~30 мс в тестовой конфигурации), который на основе эвристик (исходя из статистики сработавших алгоритмов, настроек администратора) определяет наиболее приемлемый подход – проверка исключительно на клиенте (с помощью ограниченной проверки), проверка на сервере и каким именно алгоритмом детектировать (по табл. 2 хорошо видно, что на малых словарях используется точный поиск, на средних и больших – фонетический, сервер и метод шинглов оказался не задействован ни в одном из примеров).

ЗАКЛЮЧЕНИЕ

В автономном режиме система может применяться для блокирования нежелательного контента в учебных учреждениях, на домашних компьютерах пользователей, быть составной частью поисковых систем, родительского контроля. Систему можно использовать для формирования детального отчета о содержании определенного контента веб-сайта, переписки пользователя в социальных сетях и т. д.

БЛАГОДАРНОСТИ: Работа выполнена при финансовой поддержке РФФИ (грант № 15 17 08378).

Поступила в редакцию 29 августа 2017 г.

Лопатин Дмитрий Валерьевич, Тамбовский государственный университет им. Г.Р. Державина, г. Тамбов, Российская Федерация, кандидат физико-математических наук, доцент, директор Центра компьютерной безопасности, e-mail: +79107540080@ya.ru

Чиркин Евгений Сергеевич, Тамбовский государственный университет им. Г.Р. Державина, г. Тамбов, Российская Федерация, инженер Центра компьютерной безопасности, e-mail: ccs.tmb@ya.ru

UDC 004.853
DOI: 10.20310/1810-0198-2017-22-5-1133-1137

SAFE CONTENT CREATION SYSTEM

© D.V. Lopatin, E.S. Chirkin
Tambov State University named after G.R. Derzhavin
33 Internatsionalnaya St., Tambov, Russian Federation, 392000
E-mail: ccs.tmb@ya.ru

The safe content creation system for different categories of users is described. For reliable detection of negative content, we have identified the entities of the corps of negative content from a large array of texts. The system is built on a client-server architecture and includes a client part in the form of a browser extension. The server module has the function of loading user content for later retrieval based on proven methods (naive search, regular expression search, shingles with a number of modifications, phonetic search) or a combination

of methods. For the user, the system modifies the unwanted content of full or partial blocking (replacement) of the content of the web page.
Keywords: system; content; security; user; parental control

ACKNOWLEDGEMENTS: The work is fulfilled under financial support of Russian Foundation for Basic Research (Project no. 15 17 08378).

Received 29 August 2017

Lopatin Dmitrii Valerevich, Tambov State University named after G.R. Derzhavin, Tambov, Russian Federation, Candidate of Physics and Mathematics, Associate Professor, Director of Center for Computer Security, e-mail: 79107540080@ya.ru

Chirkin Evgeniy Sergeevich, Tambov State University named after G.R. Derzhavin, Tambov, Russian Federation, Engineer of Computer Security Center, e-mail: ccs.tmb@ya.ru

Для цитирования: *Лопатин Д.В., Чиркин Е.С.* Система формирования безопасного контента // Вестник Тамбовского университета. Серия Естественные и технические науки. Тамбов, 2017. Т. 22. Вып. 5. С. 1133-1137. DOI: 10.20310/1810-0198-2017-22-5-1133-1137

For citation: Lopatin D.V., Chirkin E.S. Sistema formirovaniya bezopasnogo kontenta [Safe content creation system]. *Vestnik Tambovskogo universiteta. Seriya Estestvennye i tekhnicheskie nauki – Tambov University Reports. Series: Natural and Technical Sciences*, 2017, vol. 22, no. 5, pp. 1133-1137. DOI: 10.20310/1810-0198-2017-22-5-1133-1137 (In Russian, Abstr. in Engl.).