УДК 004.89
doi:10.15217/issn1684-8853.2016.6.65

# УЛУЧШЕНИЕ КАТЕГОРИРОВАНИЯ ВЕБ-САЙТОВ ДЛЯ БЛОКИРОВКИ НЕПРИЕМЛЕМОГО СОДЕРЖИМОГО НА ОСНОВЕ АНАЛИЗА СТАТИСТИКИ HTML-ТЭГОВ

***Д. А. Новожилов***[а, б], *студент*
***А. А. Чечулин***[а], *канд. техн. наук, старший научный сотрудник*
***И. В. Котенко***[а], *доктор техн. наук, профессор*
[а]*Санкт-Петербургский институт информатики и автоматизации РАН, Санкт-Петербург, РФ*
[б]*Санкт-Петербургский государственный электротехнический университет «ЛЭТИ», Санкт-Петербург, РФ*

***Постановка проблемы:*** *постоянный рост объема доступной информации в сети Интернет приводит к повышению сложности обнаружения нежелательной и вредоносной информации. Существующие системы используют автоматическую классификацию по текстовому содержимому веб-сайтов, однако данный метод не подходит для веб-сайтов с изменчивым содержимым, таких как новости, форумы и т. п.* ***Цель исследования:*** *повысить защищенность пользователей от нежелательной информации за счет улучшения качества категорирования веб-сайтов методами Data Mining для автоматизированных систем родительского контроля.* ***Результаты:*** *разработаны улучшенные алгоритмы классификации веб-сайтов и прототип системы родительского контроля, который осуществляет классификацию веб-сайтов, используя их структурные особенности. Основная идея заключается в анализе не текстовых признаков, а статистики HTML-тэгов, которая представляет собой совокупность их частот встречаемости (отношение числа экземпляров данного тэга к общему количеству тэгов на странице, выраженного в процентах). Всего алгоритм выбирает 25 основных тэгов по всей выборке, после чего для каждого из сайтов считается его статистика. Приведена архитектура системы категорирования, состоящей из нескольких программных модулей, написанных на языке Perl, и специального программного обеспечения RapidMiner. Для разработанного прототипа проведены эксперименты на нескольких наборах данных, после чего выполнено сравнение качества категорирования при использовании текстовых, структурных признаков, а также их комбинации. Полученные результаты показали, что анализ статистики тэгов не может использоваться в качестве самостоятельного метода, но является полезным дополнением к системам, опирающимся на текстовую классификацию (позволяет повысить ее качество в метрике «аккуратность» от 6,9 до 10,6 % в зависимости от количества категорий).* ***Практическая значимость:*** *данный подход может применяться для повышения эффективности поиска информации, запрещенной законами Российской Федерации: пропаганды экстремизма; разжигания ненависти и вражды; пропаганды порнографии, наркотиков, антиобщественного поведения и т. д. Также данный подход может использоваться в системах родительского контроля для ограничения доступа к определенным видам информации по возрастным категориям.*

***Ключевые слова*** *— Data Mining, анализ данных, защита от информации, категорирование веб-сайтов, анализ HTML-тэгов, статистика тэгов.*

## Введение

Как известно, методы Data Mining занимаются обнаружением в данных скрытых знаний: неизвестных, нетривиальных и практически полезных. В ходе работ по анализу данных часто возникает необходимость отнесения исследуемого объекта к одному из множества заранее определенных классов — задача классификации. Правильное ее решение приводит к значительным успехам во многих областях. Например, применяемая в сфере коммерции персонализация клиентов, под которой понимают автоматическое распознавание их принадлежности к определенной целевой аудитории, помогает компаниям проводить более гибкую маркетинговую политику. Не менее значимо и обеспечение безопасности при операциях с пластиковыми картами в электронных платежных системах. Data Mining на основе действий клиента позволяет относить его к одной из двух категорий: «легальный пользователь» или «злоумышленник», — обнаруживая таким образом случаи мошенничества.

Наше время отмечено непрерывным развитием и повсеместным распространением Интернета. В связи с этим возрастает значение автоматических систем классификации, распределяющих веб-страницы по категориям и блокирующих те из них, которые являются нежелательными или оскорбительными. Это бывает чрезвычайно важно, например, для ограждения детей от сайтов с неприемлемым содержимым или для противодействия распространению вредоносного и пиратского контента. Именно поэтому данному вопросу уделяется все большее внимание.

Существует множество различных подходов к классификации сайтов. Среди них наиболее эффективным и широко используемым является анализ текстового содержания веб-страниц. Однако присутствуют некоторые категории сайтов: forum, blog, news, — которые почти не различаются текстовым наполнением, тогда как структурные особенности у них разные. В подобных ситуациях переходят к другим методам классификации, например, анализу URL-адресов страниц или HTML-тэгов их разметки. Один из вари-

антов при последнем подходе заключается в том, чтобы проверять наличие/отсутствие тех или иных тэгов.

В данной статье предлагается оригинальный подход, который, в отличие от уже существующих, основывается на анализе статистики HTML-тэгов, представляющей собой отношение всех вхождений того или иного тэга к общему числу тэгов на сайте.

Основная цель исследования — разработать подход к классификации веб-сайтов на основе анализа их структурных особенностей, чтобы повысить качество категорирования в тех случаях, когда классификация по тексту затруднена.

## Обзор существующих решений

У задачи классификации сайтов существуют различные методы решения на основе анализа 1) текстового содержимого веб-страниц, 2) URL, 3) HTML-тэгов.

Наиболее широко применяется метод классификации по тексту, состоящий из двух последовательных этапов. На первом проводится подготовка данных с переводом их в форму, воспринимаемую классификатором. Один из примеров последовательности действий на данном этапе — удаление тэгов разметки и извлечение текстового содержимого веб-страниц, выполнение операции стемминга (сохранение основы слов и отбрасывание их окончаний), исключение знаков препинания, стоп-слов в виде предлогов, союзов, местоимений и т. д. Второй этап состоит в подаче предварительно обработанных данных на тот или иной классификатор (Naive Bayes, SVM и т. д.). Чаще используются методы с разделением на тестовую и обучающую выборки (supervised method). Наглядным примером может быть статья [1] с описанием метода SVM. Однако в работе [2] предлагается метод без предварительного обучения (unsupervised method), предназначенный для классификации по тексту с небольшими затратами ресурсов или для создания обучающих выборок. По нему документ делится на предложения, а затем каждому предложению сопоставляется категория на основе предварительно подготовленных списков ключевых слов и метрики подобия предложений (sentence similarity measure).

Интересный пример различных вариантов текстовой классификации приведен в статье [3], в которой в ходе рассмотрения техник определения спама предлагаются варианты категоризации на основе общего числа слов на странице, средней длины слова, принадлежности слов веб-страницы к набору из наиболее часто встречаемых слов, вычисления статистики $n$-грамм (комбинаций из $n$ символов).

Другая альтернатива — перейти от рассмотрения документов как наборов слов к анализу их значений, которые берутся из лексических баз данных. Это имеет смысл, поскольку, например, в русском языке слово «коса» может указывать на заплетенные волосы, садовый инструмент или каменную гряду. Аналогично английское слово «base» может означать военный лагерь или термин в бейсболе. Однако эксперименты показали, что рассмотрение смысла слов хотя и несколько повышает величину аккуратности, но не ведет к значительному улучшению категоризации [4].

Минус текстовой классификации состоит в том, что она не учитывает особенности веб-страниц: HTML-документ связан ссылками с другими документами, содержит изображения и иные нетекстовые элементы. Также трудности вызывают категории, обладающие сходным текстовым наполнением, но различающиеся по своей структуре (например, blog, forum, chat).

По изложенным причинам получил развитие метод, основанный на анализе URL. Здесь исходят из предположения, что страницу в Интернете будут редко посещать, если она не вызывает интерес у возможных читателей. То есть адрес сайта должен каким-то образом отражать его тематику [5]. Один из способов анализа заключается в разбиении URL на составные части, которые и будут анализироваться. Такой подход реализован при анализе URL в целях защиты от фишинговых сайтов [6]. Также имеет значение, на какой позиции находится тот или иной фрагмент адреса сайта. Например, авторы приводят следующие ссылки, содержащие фрагмент «paypal» и иллюстрирующие эту мысль: http://www.paypal.com/ и http://www.paypal.com.hostingcompany.com/.

Таким образом, каждый фрагмент URL представляется в виде двумерного вектора, содержащего сам фрагмент и его позицию, которые затем подаются на вход обученному классификатору.

Другой способ состоит в использовании длины имени хоста и всего URL, подсчете количества в нем различных символов (например, «.») и анализе заключенных между этими символами фрагментов URL. Кроме того, используются также признаки на основе информации о хосте (географические особенности, дата регистрации, величина TTL и т. д.). Все эти атрибуты подаются на вход какому-либо классификатору (Naïve Bayes, SVM, Logistic Regression) [7].

Дальнейшее разделение URL на фрагменты может быть проведено, в частности, с использованием энтропии, что позволяет разбивать на составные части названия доменов, в которых несколько слов слиты воедино, например «activatealert». То из пробных разбиений, которое имеет наименьшую энтропию среди остальных, станет наиболее вероятным новым фраг-

ментом [8]. В работах [8, 9] упоминается также о способе, связанном с анализом последовательности *n*-грамм (комбинаций из *n* символов), для которых считается частота встречаемости.

Метод на основе *n*-грамм способен показывать хорошие результаты категоризации при решении частных задач (спам/обычное письмо, phishing/benign), однако в общем случае, при произвольном количестве и составе категорий, качество классификации снижается. Главная причина заключается в том, что в действительности не всегда адрес страницы в Интернете совпадает с ее содержанием.

Таким образом, для выявления категорий, основанных на структурных признаках, необходимо искать другие методы, одним из которых может быть использование информации о HTML-тэгах сайта. Здесь также существуют различные подходы к анализу.

Важным источником может служить информация, заключенная в таких тэгах, как <title> или <meta>, которая, наряду с текстовым содержимым веб-страниц, извлекалась специальным парсером [10–12].

С другой стороны, существуют методы, основанные на подсчете количества тэгов на странице [12, 13].

В данной статье рассмотрен оригинальный подход, который, в отличие от существующих, основывается на анализе не содержания или количества HTML-тэгов на странице, а их статистики, которая определяется как отношение всех вхождений того или иного тэга к общему числу тэгов на сайте. Настоящая работа является продолжением исследований в области защиты от нежелательной информации, проводимых лабораторией и изложенных в ряде статей [14–18].

## Предлагаемый подход

Как известно, веб-страницы отличаются от обычных документов прежде всего тем, что они полуструктурированы (semi-structured) с помощью HTML-тэгов разметки, связаны между собой ссылками, содержат фрагменты кода, исполняемого как на стороне сервера, так и у клиента. Поэтому не обязательно ограничиваться исключительно текстовой классификацией, можно воспользоваться другими методами, более полно учитывающими специфику анализируемых данных. Одно из возможных решений — применение других подходов, связанных со структурными особенностями веб-страниц, например, с анализом HTML-тэгов.

Предлагаемый метод также не опирается на сохраненные исходные тексты веб-страниц для последующего их анализа, а работает со статистикой HTML-тэгов. Под статистикой $S$ тэгов понимается совокупность их частот встречаемости $f_i$, которые считаются как отношение числа экземпляров данного тэга $n_i$ к общему количеству тэгов на странице $N$, выраженное в процентах. Результат округляется до десятых для обеспечения большей информативности:

$$S=\cup f_i;\ f_i=(n_i/N)\cdot 100\%.$$

Следует отметить, что такое решение было найдено не сразу, и сначала анализировалось простое количество тэгов каждого вида на странице. Однако данный подход является не совсем правильным, поскольку, например, 100 тэгов  на страницах, состоящих из 250 и 1000 тэгов, некорректно сравнивать, и они указывают на совершенно разный результат.

Итоговый классификатор строится на основе алгоритмов Naïve Bayes и Decision Tree, базовые предсказания которых объединяются на верхнем уровне с помощью Stacking. Далее приводится более подробное описание перечисленных методов.

Алгоритм Naïve Bayes основан на применении теоремы Байеса, известной из теории вероятности. Отличительной чертой является «наивное» предположение о независимости событий. В применении к задаче классификации данных она может формулироваться следующим образом.

Имеется множество $W=\{w_1, w_2, ..., w_n\}$ веб-страниц для категоризации и задано множество категорий $C=\{c_1, c_2, ..., c_k\}$. Если каждая веб-страница может быть отнесена только к одной категории, тогда вероятность для сайта $s$ попасть в категорию $c$ определяется формулой Байеса

$$p=\arg\max_{c\in C} P(c|S)=$$
$$=\arg\max_{c\in C}(P(c)\cdot P(S|C)/P(s)).$$

Вероятность $P(s)$ может не учитываться, поскольку она является постоянной величиной (не зависит от категории).

Преимуществом алгоритма является малое количество данных для обучения, необходимых для оценки параметров при классификации, простота реализации и низкие вычислительные затраты. В тех редких случаях, когда признаки действительно независимы (или почти независимы), наивный байесовский классификатор (почти) оптимален.

Среди недостатков — относительно низкое качество классификации в большинстве реальных задач, где нарушается предположение независимости (например, в естественном языке вероятность появления слова сильно зависит от контекста).

Деревья принятия решений (Decision Tree) — иерархическая структура данных, которая не только способна решить задачу классификации, но и позволяет с легкостью интерпретировать полученные результаты, чтобы объяснить, почему

объект был классифицирован тем или иным образом [19].

В деревьях решений определяется значение целевого атрибута, обозначенного как label. Основа генерации деревьев решений — рекурсивное разбиение выборки на подмножества на основе значений атрибутов, выбираемых по определенному критерию. Рекурсия прекращается, когда все элементы подмножества или большинство имеют одинаковое значение атрибута label. Возможны и другие варианты остановки алгоритма, например достижением максимальной глубины дерева.

К преимуществам метода относятся быстрый процесс обучения; генерация правил в областях, где эксперту трудно формализовать свои знания; извлечение правил на естественном языке и простота интерпретации.

Среди недостатков — неприменимость к наборам данных, где число возможных исходов велико. Тогда деревья «переполнены данными», имеют много узлов и ветвей, и в них очень трудно разобраться. Для деревьев решений обычно характерны высокие результаты по метрике точности при снижении показателя полноты. Кроме того, одна из особенностей деревьев решений заключается в том, что они нередко могут образовывать «поглощающую» категорию, к которой ошибочно относится большое количество данных из других категорий.

Для повышения общего качества классификации можно использовать не только какой-либо один алгоритм анализа данных, но и их комбинации, получившие названия мета-алгоритмов. При таком подходе обучается набор базовых классификаторов, после чего результаты их прогнозов объединяются, например, путем взвешенного усреднения или голосования.

Существуют различные мета-алгоритмы. Один из них — метод под названием Boosting, особенность которого состоит в том, что базовые классификаторы обучаются последовательно. При этом обучающий набор данных для каждого последующего базового классификатора зависит от точности прогноза предыдущего. Алгоритм Bagging предлагает другой подход, в котором из всех собранных данных случайным образом выбираются подмножества (случайный выбор с возвратом), которые подаются на вход каждой из моделей мета-алгоритма, а их результаты комбинируются. Существуют и другие методы.

В данной работе применяется алгоритм Stacking. Он использует в качестве базовых моделей различные классификационные алгоритмы, обучаемые на одинаковых данных. Затем метаклассификатор обучается на исходных данных, дополненных результатами прогноза базовых алгоритмов. Идея Stacking заключается в том, что мета-алгоритм учится различать, какому из базовых алгоритмов следует «доверять» на каких областях входных данных.

Для оценки качества классификации используются такие метрики, как точность (precision), полнота (recall), аккуратность (accuracy) и F-мера — метрика, объединяющая в себе информацию о точности и полноте, представляющая собой гармоническое среднее между этими двумя метриками.

Следует отметить, что для класса систем, выполняющих функции родительского контроля или защиты от вредоносного программного обеспечения, которые и находятся в фокусе данного исследования, особое значение приобретает метрика «точность», поскольку большое количество ложных срабатываний может послужить причиной отказа от использования подобных систем.

## Реализация предлагаемого подхода

Задачу нахождения частоты тэгов можно решать несколькими различными способами: 1) поиском тэгов по всему HTML-документу и подсчетом количества вхождений каждого из них; 2) используя представление HTML-документа в виде дерева, которое значительно упрощает решение задачи, предоставляя различные функции навигации и доступа к его элементам.

Одним из аргументов в пользу второго подхода явилось наличие подобной древовидной структуры данных — она строилась для нужд анализа документа и сохранения его текстового содержимого в файл без HTML-тэгов (рис. 1, *а–в*).

Таким образом, модуль, отвечающий за статистику тэгов на странице, был построен на основе модификации уже имеющегося модуля определения основного языка веб-страницы и сбора данных, что значительно облегчило его разработку. Обе программы написаны на языке Perl. Этот язык программирования включает мощные инструменты обработки текста, которые делают его идеальным для работы с HTML, XML и другими языками разметки или естественными.

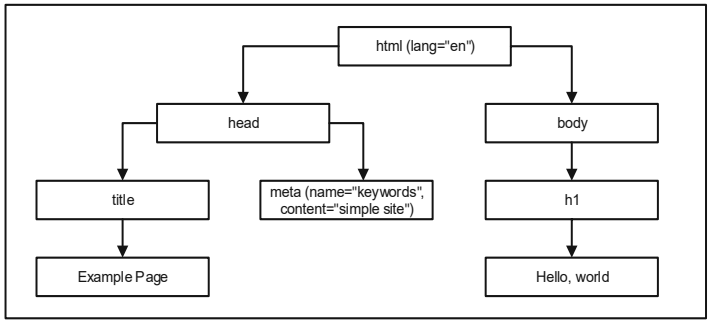
При использовании функционала стандартных классов HTML::TreeBuilder и HTML::Element была создана функция, которая выгружает все тэги узла и его потомков в ассоциативный массив (рис. 2).

Ключом является название тэга, а значениями — все его представители, количество которых подсчитывается. Если функцию применить к корню дерева, то будет получено решение задачи. Для дальнейшего анализа используются все тэги, частота которых превышает 2 % (установленное опытным путем значение, позволяющее исключить из рассмотрения общие для всех страниц тэги, такие как <html>, <title>, <head>,

*а)*

```
<html lang="en">
<head>
        <title>Example page</title>
        <meta name="keywords" content="simple site"
</head>
<body>
        <h1>Hello, world</h1>
</body>
</html>
```

*б)*




*в)*

```
element#1:      _tag:'html'
                _parent: none
                _content: [element#2, element#5]
                lang='en'

element#2:      _tag:'head'
                _parent: element#1
                _content: [element#3, element#4]

element#3:      _tag:'title'
                _parent: element#2
                _content: [text segment "Example Page"]

element#4:      _tag:'meta'
                _parent: element#2
                _content: none
                name = keywords
                content = simple site

element#5:      _tag:'body'
                _parent: element#1
                _content: [element#6]

element#6:      _tag:'h1'
                _parent: element#5
                _content: [text segment "Hello, world"]
```

■ ***Рис. 1.*** Переход от простейшего HTML-исходника (*а*) к его модели в виде дерева (*б*) и программному представлению (*в*) в модуле HTML::Element

```
{
        'br' => [...list of all <br> elements...];

        'div' => [...list of all <div> elements...];

        'li' => [...list of all <li> elements...].
}
```

■ ***Рис. 2.*** Структура ассоциативного массива в модуле HTML::Element

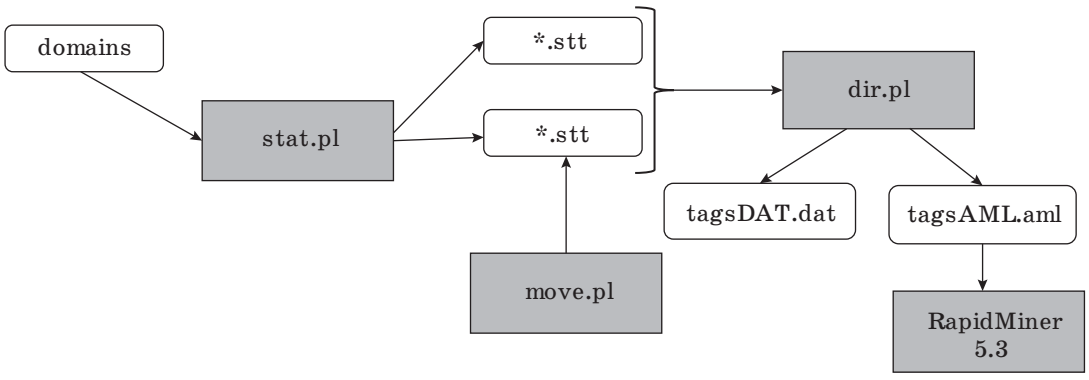
<body> и т. д., чтобы не добавлять их в стоп-слова).

Архитектура созданной системы представлена на рис. 3. В ней в файле domains содержатся ссылки на сайты, объединенные определенной тематикой (относящиеся к одной категории классификации). Модуль считывает адреса веб-страниц, переходит по каждой из ссылок, и вся статистика тэгов по каждому сайту сохраняется в отдельный файл со специфическим расширением *.stt, позволяющим отличать его от других.

В соответствии с данной архитектурой набор *.stt-файлов попадает на обработку в модуль dir.pl, который из собранных данных выбирает Top-25 тэгов, в разной степени встречающихся на каждом из сайтов в имеющейся выборке.

Так формируется своеобразный базис, по которому раскладывается статистика тэгов каждой конкретной страницы, а соответствующие частоты будут коэффициентами.

Результатом работы модуля являются файлы «tagsDAT.dat» и «tagsAML.aml», необходимые на последующих этапах в программном обеспечении анализа данных RapidMiner 5.3 [20], в котором в виде блоков реализованы все используемые в настоящем исследовании классификаторы.




■ ***Рис. 3.*** Архитектура системы классификации на основе анализа статистики HTML-тэгов

AML-файлы содержит xml-подобный текст с заголовками, а их значения содержатся в DAT-файле. Модуль move.pl служит для перемещения файлов.

Предполагается, что данный подход позволит лучше различать категории, которые обычно путаются при классификации по тексту из-за одинакового смыслового содержания: chat — forum, guns — hunting и т. д.

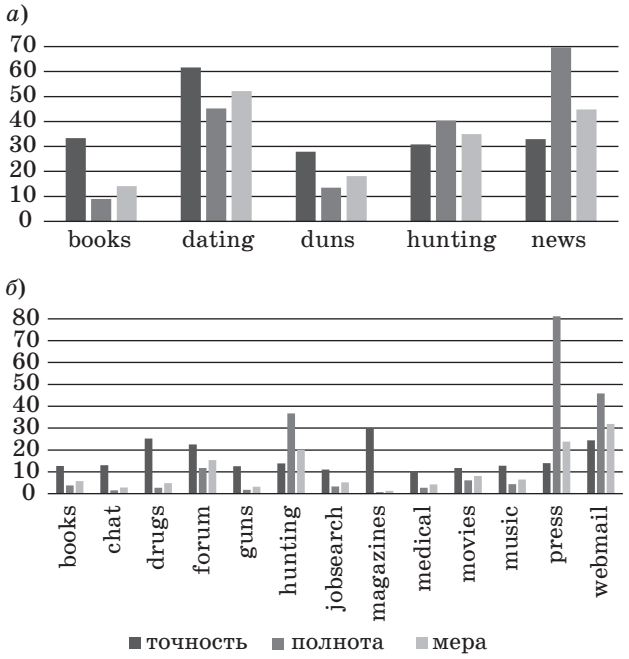
### Результаты экспериментов

Эксперименты проводились на двух наборах данных («set1» и «set2»). Набор «set1» создан на основе исходных данных (файлов domains) с сайта URLBlacklist.com [21], в которые вошли категории «books, hunting, news, dating, guns», для каждой из которых отобрано по 1000 сайтов. Набор «set2» включает в себя следующие источники данных: материалы сайта URLBlacklist.com [21], объединенные с частью категорий, взятых из списков «Shalla Secure Services KG» [22]. Два различных каталога было использовано для того, чтобы выделить общие черты сайтов, а также по причине нехватки исходных данных по определенным категориям в одном источнике и достаточного их количества в другом. Окончательно было выбрано 13 категорий: «books, chat, drugs, forum, guns, hunting, jobsearch, magazines, medical, movies, music, press, webmail», — в каждой из которых около 1500 сайтов.

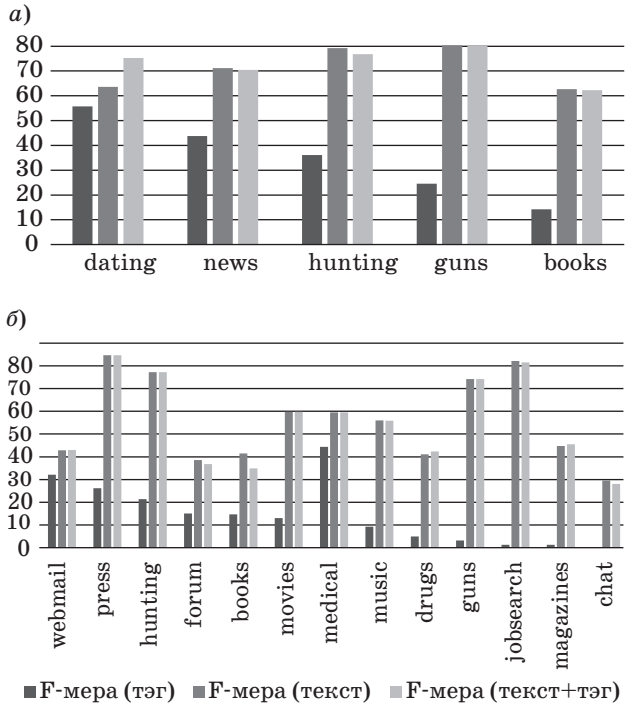
При подготовке исходных данных уделялось внимание границам категорий. Такие неоднородные категории, как «radio-tv» или «audio-video», исключались из рассмотрения, поскольку фактически каждая из них подразделяется на две. Категории drugs и medical, guns и hunting, наоборот, были взяты специально, чтобы оценить работу в случаях, когда некоторые специфичные слова и сочетания могут быть для них общими.

Результаты экспериментов для первого набора представлены на рис. 4, *а*. Сравнение результатов классификации по тэговым и текстовым признакам для первого набора приведено на рис. 5, *а*. Результаты экспериментов для второго набора представлены на рис. 4, *б*. Сравнение результатов классификации по тэговым и текстовым признакам для второго набора приведено на рис. 5, *б*.

По анализу результатов экспериментов можно сделать вывод о невысоком в целом качестве, не позволяющем использовать данный метод как основной инструмент классификации. Набор «set1» дает более высокое значение аккуратности, равное 35,43 %, поскольку содержит меньше «спорных категорий», между которыми возможно пересечение (только guns и hunting). Для набора «set2» аккуратность снижается до 15,08 %, при этом отчетливо видно, что категории press и hunting стали «поглощающими». Эта особенность деревьев решений уже отмечалась выше — к ним верно относятся сайты, действительно входящие в эти категории, но также многие веб-страницы, не являющиеся таковыми.




■ ***Рис. 4.*** Значения основных метрик для набора «set1» (*а*) и «set2» (*б*)




■ ***Рис. 5.*** Категории набора «set1» (*а*) и «set2» (*б*), упорядоченные по убыванию F-меры для классификации по тэгам

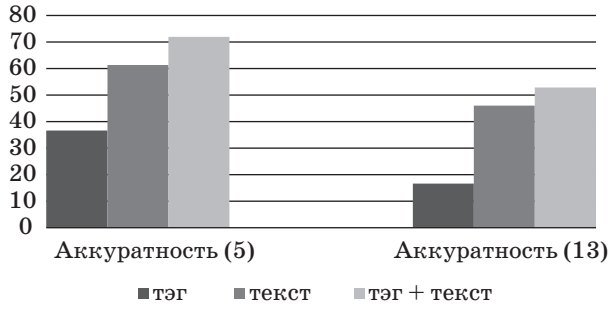



***Рис. 6.*** Показатель аккуратности для подходов, основанных на разных признаках и их комбинации

Из рассмотрения результатов при классификации совместно по тэгам и тексту можно заключить, что каждый из подходов позволяет выделить разные категории по критерию максимума F-меры. На наборе «set1» для текста — это hunting и guns, незначительно отстоящие друг от друга по величине F-меры, что отражает одну из проблем при классификации: на охоте используется оружие, а сайты, торгующие оружием, могут предлагать его для охоты. При использовании статистики тэгов первые места занимают уже другие категории — dating и news, а различие между hunting и guns несколько увеличивается. Набор «set2» демонстрирует похожую картину: в лидерах по тексту — press, jobsearch, guns и hunting, слабо отличающиеся друг от друга. Применение статистики тэгов выводит на первый план webmail и press и позволяет улучшить категоризацию для guns и hunting за счет более сильного различия между ними.

Значения аккуратности для подходов, основанных на анализе тэгов, текста и их комбинации, приведены на рис. 6. Результаты отражают повышение качества классификации при объединении данных подходов как для пяти, так и для 13 категорий.

Таким образом, проведенные исследования показывают, что предложенный подход на основе статистики HTML-тэгов самостоятельно не решает задачи категорирования, но может быть хорошим дополнением к текстовой классификации при выделении категорий, отличающихся по структурным особенностям.

## Заключение

В данной статье рассматриваются подходы к категоризации веб-страниц, не обладающих существенными отличиями при текстовой классификации, но имеющих различную структуру. В основе предлагаемого метода лежит использование статистики HTML-тэгов, которая подается на вход классификаторов.

Проведен полный цикл исследований. Эксперименты показали следующие значения аккуратности: 35,43 % для набора «set1» и 13,69 % для набора «set2». Выполнены сравнение результатов классификации на основе тэгов с текстовой классификацией на основе F-меры. Проанализирована комбинированная схема классификации, использующая текстовые и структурные признаки сайта одновременно. Так, для пяти категорий добавление анализа тэгов повысило аккуратность на 10,6 %, а для 13 категорий — на 6,9 %. Полученные результаты свидетельствуют о том, что уровень классификации по тэгам недостаточен для того, чтобы применять данный метод в качестве самостоятельного, однако он может быть использован как полезное дополнение к существующим системам с текстовой классификацией. Исследованные принципы могут применяться для улучшения качества систем защиты от информации, таких как системы родительского контроля. К дальнейшим направлениям исследований можно отнести использование сайта DMOZ.org в качестве источника исходных данных, поскольку применяемые на текущий момент каталоги интернет-ресурсов не обладают достаточным их количеством. Еще одна задача, стоящая на данном этапе, — поиск других классификаторов и их комбинаций, что позволит объединить анализ данных по тексту и статистике тэгов, избавиться от характерных для деревьев решений «поглощающих категорий». Также важна модификация имеющихся схем с целью обеспечить распараллеливание и ускорение анализа данных.

Работа выполнена при финансовой поддержке РФФИ (проекты № 14-07-00697, 14-07-00417, 15-07-07451, 16-37-00338), Российского научного фонда (проект № 15-11-30029) и при частичной поддержке бюджетных тем № 0073-2015-0004 и 0073-2015-0007.

## Литература

1. **Joachims T.** Text Categorization with Support Vector Machines: Learning with Many Relevant Features // Proc. of 10th European Conf. on Machine Learning (ECML-98), Chemnitz, Germany, April 21–23, 1998. P. 137–142.
2. **Ko Y., Seo J.** Automatic Text Categorization by Unsupervised Learning // Proc. of the 18th Conf. on Computational Linguistics (Coling-2000). 2000. P. 453–459.
3. **Ntoulas A.,** et al. Detecting Spam Web Pages through Content Analysis/ A. Ntoulas, M. Najork, M. Manasse, D. Fetterly // Proc. of the 15th Intern. World Wide Web Conf. (WWW-2006). 2006. P. 83–92.

4. **Kehagias A.,** et al. A Comparison of Word- and Sense-based Text Categorization Using Several Classification Algorithms/ A. Kehagias, V. Petridis, V. G. Kaburlasos, P. Fragkou // Journal of Intelligent Information Systems. 2000. Vol. 21(3). P. 227–247.
5. **Attardi G., Gulli A., Sebastiani F.** Automatic Web Page Categorization by Link and Context Analysis // Proc. of 1st European Symp. on Telematics, Hypermedia and Artificial Intelligence (THAI-1999). 1999. P. 105–119.
6. **Khonji M., Iraqi Y., Jones A.** Enhancing Phishing E-Mail Classifiers: A Lexical URL Analysis Approach // Intern. Journal for Information Security Research. 2012. Iss. 6. P. 236–245.
7. **Ma J.,** et al. Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs/ J. Ma, L. K. Saul, S. Savage, G. M. Voelker // Proc. of Conf. on Knowledge Discovery and Data Mining. 2009. P. 1245–1254.
8. **Kan M.-Y., Thi H. O. N.** Fast Webpage Classification Using URL Features // Proc. of Conf. on Information and Knowledge Management. 2005. P. 325–326.
9. **Geide M.** N-gram Character Sequence Analysis of Benign vs. Malicious Domains/URLs. http://analysis-manifold.com/ngram_whitepaper.pdf (дата обращения: 24.03.2016).
10. **Patil A. S., Pawar B. V.** Automated Classification of Web Sites Using Naive Bayesian Algorithm // Proc. of the Intern. Multiconf. of Engineers and Computer Scientists. 2012. P. 466.
11. **Riboni D.** Feature Selection for Web Page Classification // Proc. of the Workshop on Web Content Mapping: A Challenge to ICT (EURASIA-ICT). 2002. P. 121–128.
12. **Kotenko I.,** et al. Analysis and Evaluation of Web Pages Classification Techniques for Inappropriate Content Blocking/ I. Kotenko, A. Chechulin, A. Shorov, D. Komashinsky // Proc. of 14th Industrial Conf. on Data Mining (ICDM 2014). 2014. P. 39–54.
13. **Meshkizadeh S., Masoud-Rahmani A.** Webpage Classification Based on Compound of Using HTML Features & URL Features and Features of Sibling Pages // Intern. Journal of Advanced Computer Technology. 2010. Iss. 2(4). P. 36–46.
14. **Novozhilov D., Kotenko I., Chechulin A.** Improving the Categorization of Web Sites by Analysis of Html-Tags Statistics to Block Inappropriate Content // Proc. of the 9th Intern. Symp. on Intelligent Distributed Computing (IDC-2015), Guimaraes, Portugal, October 7–9, 2015. 2016. P. 257–263. doi:10.1007/978-3-319-25017-5_24
15. **Kotenko I., Chechulin A., Komashinsky D.** Evaluation of Text Classification Techniques for Inappropriate Web Content Blocking // Proc. of the IEEE 8th Intern. Conf. on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS-2015), Warsaw, Poland, Sept. 24–26, 2015. 2015. P. 412–417.
16. **Котенко И. В., Чечулин А. А., Комашинский Д. В.** Автоматизированное категорирование веб-сайтов для блокировки веб-страниц с неприемлемым содержимым // Проблемы информационной безопасности. Компьютерные системы. 2015. № 2. С. 69–79.
17. **Комашинский Д. В., Котенко И. В., Чечулин А. А.** Категорирование веб-сайтов для блокирования веб-страниц с неприемлемым содержимым // Системы высокой доступности. 2011. № 2. С. 102–106.
18. **Комашинский Д. В.** и др. Автоматизированная система категорирования веб-сайтов для блокирования веб-страниц с неприемлемым содержимым/ Д. В. Комашинский, И. В. Котенко, А. А. Чечулин, А. В. Шоров // Системы высокой доступности. 2013. № 3 (9). С. 119–127.
19. RapidMiner Operator Reference Guide. http://docs.rapidminer.com/studio/operators/ (дата обращения: 24.03.2016).
20. RapidMiner 5.3. http://rapidminer.com/ (дата обращения: 24.03.2016).
21. URLBlacklist. http://urlblacklist.com/ (дата обращения: 24.03.2016).
22. Shalla Secure Services KG. http://www.shallalist.de/ (дата обращения: 24.03.2016).

UDC 004.89
doi:10.15217/issn1684-8853.2016.6.65

**Improving Website Categorization Based on HTML Tag Statistics for Blocking Unwanted Content**

Novozhilov D. A.[a,b], Student, novozhilov@comsec.spb.ru
Chechulin A. A.[a], PhD, Tech., Senior Researcher, chechulin@comsec.spb.ru
Kotenko I. V.[a], Dr. Sc., Tech., Professor, ivkote@comsec.spb.ru
[a]Saint-Petersburg Institute for Informatics and Automation of the RAS, 39, 14 Line, V. O., 199178, Saint-Petersburg, Russian Federation
[b]Saint-Petersburg State Electrotechnical University «LETI», 5, Prof. Popov St., 197376, Saint-Petersburg, Russian Federation

**Introduction:** The continuous development and ubiquity of the Internet lead to a higher complexity of detecting unwanted and malicious information. The existing systems usually use automatic classification by textual content of websites, but this approach cannot be applied to websites with changeable content like news, forums, etc. **Purpose:** The goal is to enhance the protection against unwanted

or inappropriate information through improving the categorization quality by using Data Mining techniques for automated parental control systems. **Results:** Improved algorithms have been developed for website classification, along with a prototype of a parental control system. The novelty of the proposed approach is using not the textual content but the statistics of HTML tags (the ratio of the number of occurrences of a certain tag on a page to the total number of all tags on this page). The algorithm selects 25 main tags from a set of websites and then calculates tags' statistics for each website. The paper also describes the architecture of the categorization system which consists of several Perl modules and special RapidMiner software. For the developed prototype, some experiments on preformed datasets were carried out, with the comparison of categorization quality between text, structure features and their combinations. The results showed that the analysis of tag statistics is not sufficient to replace all the other methods. But it can be a useful complement to the existing systems with textual classification, able to increase their quality from 6.9 to 10.6% in accuracy metrics, depending on the number of categories. **Practical relevance:** This approach can be used to improve the efficiency of search for information forbidden by the laws of the Russian Federation (propaganda of extremism, pornography, drugs, anti-social behavior, etc). Also, this approach can be used in parental control systems to deny access to certain types of information according to age categories.

**Keywords** — Data Mining, Data Analysis, Protection from Information, Website Categorization, HTML Tag Analysis, Tag Statistics.

**References**

1. Joachims T. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *Proc. of 10th European Conf. on Machine Learning (ECML-98)*, Chemnitz, Germany, April 21–23, 1998, pp. 137–142.
2. Ko Y., Seo J. Automatic Text Categorization by Unsupervised Learning. *Proc. of the 18th Conf. on Computational linguistics (Coling-2000)*, 2000, pp. 453–459.
3. Ntoulas A., Najork M., Manasse M., Fetterly D. Detecting Spam Web Pages through Content Analysis. *Proc. of the 15th Intern. World Wide Web Conf. (WWW-2006)*, 2006, pp. 83–92.
4. Kehagias A., Petridis V., Kaburlasos V. G., Fragkou P. A Comparison of Word- and Sense-based Text Categorization Using Several Classification Algorithms. *Journal of Intelligent Information Systems*, 2000, vol. 21(3), pp. 227–247.
5. Attardi G., Gulli A., Sebastiani F. Automatic Web Page Categorization by Link and Context Analysis. *Proc. of 1st European Symposium on Telematics, Hypermedia and Artificial Intelligence (THAI-1999)*, 1999, pp. 105–119.
6. Khonji M., Iraqi Y., Jones A. Enhancing Phishing E-Mail Classifiers: A Lexical URL Analysis Approach. *Intern. Journal for Information Security Research*, 2012, iss. 6, pp. 236–245.
7. Ma J., Saul L. K., Savage S., Voelker G. M. Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs. *Proc. of Conf. on Knowledge Discovery and Data Mining*, 2009, pp. 1245–1254.
8. Kan M.-Y., Thi H. O. N. Fast Webpage Classification Using URL Features. *Proc. of Conf. on Information and Knowledge Management*, 2005, pp. 325–326.
9. Geide M. *N-gram Character Sequence Analysis of Benign vs. Malicious Domains/URLs*. Available at: http://analysis-manifold.com/ngram_whitepaper.pdf. (accessed 24 March 2016).
10. Patil A. S., Pawar B. V. Automated Classification of Web Sites Using Naive Bayesian Algorithm. *Proc. of the Intern. Multiconf. of Engineers and Computer Scientists*, 2012, p. 466.
11. Riboni D. Feature Selection for Web Page Classification. *Proc. of the Workshop on Web Content Mapping: A Challenge to ICT (EURASIA-ICT)*, 2002, pp. 121–128.
12. Kotenko I., Chechulin A., Shorov A. Komashinsky D. Analysis and Evaluation of Web Pages Classification Techniques for Inappropriate Content Blocking. *Proc. of 14th Industrial Conf. on Data Mining (ICDM 2014)*, 2014, pp. 39–54.
13. Meshkizadeh S., Masoud-Rahmani A. Webpage Classification Based on Compound of Using HTML Features & URL Features and Features of Sibling Pages. *Intern. Journal of Advanced Computer Technology*, 2010, iss. 2(4), pp. 36–46.
14. Novozhilov D., Kotenko I., Chechulin A. Improving the Categorization of Web Sites by Analysis of Html-Tags Statistics to Block Inappropriate Content. *Proc. of the 9th Intern. Symp. on Intelligent Distributed Computing (IDC-2015)*, Guimaraes, Portugal, October 7–9, 2015, 2016, pp. 257–263. doi:10.1007/978-3-319-25017-5_24
15. Kotenko I., Chechulin A., Komashinsky D. Evaluation of Text Classification Techniques for Inappropriate Web Content Blocking. *Proc. of the IEEE 8th Intern. Conf. on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS-2015)*, Warsaw, Poland, September 24–26, 2015, pp. 412–417.
16. Kotenko I., Chechulin A., Komashinsky D. Automated Categorization of Web-sites for Inappropriate Content Blocking. *Problemy informatsionnoi bezopasnosti. Komp'iuternye sistemy* [Problems of Information Security. Computer Systems], 2015, no. 2, pp. 69–79 (In Russian).
17. Komashinskiy D., Kotenko I., Chechulin A. Categorization of Web Sites for Inadmissible Web Pages Blocking. *Sistemy vysokoy dostupnosti* [High Availability Systems], 2011, no. 2, pp. 102–106 (In Russian).
18. Komashinskiy D., Kotenko I., Chechulin A., Shorov A. Automatic System for Categorization of Websites for Blocking Web Pages with Inappropriate Contents. *Sistemy vysokoy dostupnosti* [High Availability Systems], 2013, no. 3 (9), pp. 119–127 (In Russian).
19. *RapidMiner Operator Reference Guide*. Available at: http://docs.rapidminer.com/studio/operators/ (accessed 24 March 2016).
20. *RapidMiner 5.3*. Available at: http://rapidminer.com/ (accessed 24 March 2016).
21. *URLBlacklist*. Available at: http://urlblacklist.com/ (accessed 24 March 2016).
22. *Shalla Secure Services KG*. Available at: http://www.shallalist.de/ (accessed 24 March 2016).