Блокирование Web-сайтов с неприемлемым содержанием на основании выявления их категорий

Зозуля Ю.В., Котенко И.В.
Лаборатория проблем компьютерной безопасности, СПИИРАН
{zozulya, ivkote}@comsec.spb.ru

Представляемая работа посвящена разработке общего подхода и реализующего его начального прототипа, предназначенных для решения актуальной в настоящее время (например, для систем родительского контроля) задачи категорирования веб-сайтов для систем блокирования веб-страниц с неприемлемым содержимым. Системы блокирования предназначены для обеспечения безопасности пользователя при доступе в Интернет. Существующие решения на данном этапе не предусматривают участия пользователя в выборе тематики блокируемого содержимого.

Согласно предлагаемому подходу пользователь (например, родитель) может определить список категорий, по которым будет оцениваться каждый посещаемый веб-сайт. В соответствии с выбранными категориями странице ставится оценка с учетом информации из разных источников. Оценка сайта заносится в базу данных, а затем используется и обновляется при последующем доступе к этому или другим сайтам.

Анализируется не только содержимое сайта, но и контекстная информация, включающая в себя историю оценок данного сайта, оценки соседних сайтов, а также оценки просматриваемого сайта, доступные из внешних источников.

Веб-страница анализируется автоматическим классификатором, построенном на основании обучающей выборки для каждой неприемлемой категории. Построение классификатора условно можно разделить на три этапа: индексация веб-страницы, построение и обучение классификатора, оценка качества классификации на тестовой выборке.

Процесс индексации веб-страницы необходимо проводить, поскольку классификатор не способен интерпретировать веб-страницу непосредственно. Процесс индексации веб-страницы заключается в сопоставлении ее содержимого с компактным представлением – вектором элементов и их весовых коэффициентов. В качестве элементов вектора, представляющего веб-страницу, выделены совокупности слов из разных источников: текст на целевой веб-странице; гиперссылки; элементы соседних веб-страниц (текст гиперссылки на целевую страницу, текст вокруг гиперссылки, заголовки).

Индексация веб-страницы проходит в несколько этапов:

1. Построение вектора, представляющего веб-страницу. Заключается в первоначальной интерпретации html-тегов, лексического и морфологического анализа текстового содержимого, выделения словоформ, синтаксического анализа выделенных словоформ, удаления стоп-слов.
2. Вычисление весовых коэффициентов полученного вектора. Проводится с учетом html-тегов, а также при помощи стандартных методов вычисления веса элемента веб-документа (стандартную формулу *tf-idf*).
3. Уменьшение размерности полученного вектора. Высокая размерность вектора, представляющего веб-страницу, может привести к достаточно большой вычислительной сложности процесса классификации. Наиболее простой и эффективный способ такой фильтрации – отбор наиболее значимых элементов по результатам функции, определяющей количество документов из обучающей выборки, в которых встретился соответствующий элемент вектора. Таким образом, остаются только те элементы, которые появляются в наибольшем числе веб-страниц.

Результатом работы классификатора является множество $A = \{a_1, ..., a_{|A|}\}$, где a_i - степень принадлежности каждой посещаемой страницы к соответствующей категории из выбранных пользователем.

С учетом степени принадлежности a_i для каждой из выбранных пользователем категорий и множества весовых коэффициентов $\Omega = \{\omega_1, ..., \omega_{|\Omega|}\}$ для каждой категории, проставленных пользователем, системная оценка целевой веб-страницы вычисляется в соответствии с формулой (1).

$$G_s = \max_{a_i \in A, \omega_i \in \Omega} a_i \cdot \omega_i . \quad (1)$$

Таким образом, G_s – максимальная мера принадлежности веб-страницы к любой из запрещенных пользователем категории с учетом указанной для каждой категории значимости.

С учетом настроенных параметров влияния пользовательской оценки и истории оценки результирующая оценка рассчитывается с помощью формулы (2).

$$G = \overline{G} = \omega_u G_u + (1 - \omega_u)(\omega_b G_b + (1 - \omega_b) G_s), \quad (2)$$

где G_u и ω_u – пользовательская оценка и ее коэффициент влияния, а G_b и ω_b – интегрированная оценка, сформированная с учетом истории оценок и ее коэффициент влияния.

Интегрированная оценка формируется в соответствии с алгоритмом 1.

Алгоритм 1:

1: $G_b \leftarrow G_N$

2: for **grade** of *n=N-1* to *1* do

3: $G_b \leftarrow \omega_b \cdot G_b + (1 - \omega_b) \cdot G_n$

В алгоритме G_N – последняя оценка для данной веб-страницы, занесенная в базу данных.

С учетом определенного родителем-администратором уровня фильтрации веб-страниц и вычисленного порогового значения τ_i, принимается решение, такое, что при $G \geq \tau_i$ доступ к веб-странице запрещен, а при $G < \tau_i$ – разрешен.

Подразумеваются также различные сценарии поведения системы в случае запрета доступа к веб-странице:

1. Предполагается такая реакция браузера, что пользователь может посчитать, что он ошибся при наборе адреса веб-сайта или то, что этот веб-сайт на данный момент недоступен по различным причинам, т.е. стандартное поведение браузера, если по каким-либо причинам он не может отобразить ту или иную веб-страницу.
2. Подразумевается наличие веб-страницы-заглушки, которую открывает браузер при попытке доступа к запрещенному сайту. На такой странице может быть расположена полезная информация
3. Предлагается перейти на несколько других веб-страниц, выбранных администратором, например, поисковые системы или специализированные сайты.
4. В явном виде выдается предупреждение о том, что доступ к этому сайту был ограничен из-за его неприемлемого содержимого.

Пользователь также может включить регистрацию событий, произошедших при доступе в Интернет. При этом могут регистрироваться как положительные события (доступ разрешен), так и отрицательные.

Оценки для веб-сайтов, проставленные пользователем и выставленные системой, а также степени принадлежности сайта к определенной категории, приведенные в соответствие с форматом внутренней системной оценки данного сайта, заносятся в базу оценок системы. Сроки хранения и количество хранимых оценок настраивается пользователем.

Эксперименты позволяют говорить о перспективности разработанной модели, однако существует необходимость дополнительных исследований в этой области.

Модели и прототипы разрабатывалась целенаправленно для систем блокирования неприемлемого содержимого при доступе в Интернет. Поэтому целесообразным будет их

использование именно в этих случаях. В текущем состоянии система предназначена для пользовательского компьютера и не подразумевает использование клиент-серверных приложений и распределенной структуры вычислений. Разработка таких систем требует дополнительных исследований, при этом ожидается, что их эффективность будет выше исключительно пользовательской в силу нескольких причин. Во-первых, распределенная структура вычислений значительно повышает производительность системы. Во-вторых, такая архитектура вовлекает большее количество пользователей, что, в свою очередь, подразумевает отклик, который может повысить эффективность как классификаторов, так и системы в целом.

Работа выполнена при финансовой поддержке РФФИ (проект № 10-01-00826-а) и программы фундаментальных исследований ОНИТ РАН (проект № 3.2).