УДК 519.237+004.4:234

РАЗРАБОТКА КЛАССИФИКАЦИОННОЙ СХЕМЫ ИНФОРМАЦИОННЫХ БЛОКОВ САЙТОВ

Криницкая А. И., Мартыненко Т.В.
Донецкий Национальный Технический Университет
кафедра автоматизированных систем управления
E-mail: alesya@telenet.dn.ua

Аннотация

Криницкая А.И., Мартыненко Т.В. Разработка классификационной схемы информационных блоков сайтов. *В работе обсуждаются основные подходы к выделению основного контента web-страницы. Разработана классификационная схема информационных блоков сайтов. Определены основные проблемы существующих разработок, частично решающих поставленную задачу. Приведен предлагаемый алгорим для реализации поставленной задачи.*

Общая постановка проблемы

Высокая доступность огромного количества постоянно пополняющейся информации, а также растущая популярность веб-услуг среди всех категорий пользователей обострили проблему выделения значимой для пользователя части информации. Основная проблема заключается в том, что большинство веб-сайтов содержит множество ненужной пользователю информации на страницах — так называемый «информационный шум». К нему можно отнести навигацию, связанные ссылки, элементы дизайна, рекламу. Весь этот «информационный шум» зачастую мешает нормальному восприятию необходимой информации.

Определение информационному шуму можно дать, опираясь на понятия релевантности. Релевантность – это соответствие запроса результату. Таким образом, несоответствие запроса результату будет можно трактовать как информационный шум. «Информационный шум - это когда изобилие поступающей человеку информации делает большую её часть нерелевантной (то есть не полезным сигналом, а именно "шумом")» [4].

Таким образом, будем понимать под информационным шумом ненужную, несвоевременную информацию, мешающую потребителю воспринимать другую - соответствующую его запросам.

Зачастую при визуальной фильтрации контента и оценке его значимости пользователь теряет массу времени. Для решения этой проблемы необходимо применять очистку веб-страниц от информационного шума. Обозначим несколько областей, для которых можно будет применить задачу очистки веб-страниц:

- сервисы доставки контента, когда другие способы по каким-то причинам не подходят (например, RSS лента отсутствует);
- системы по сбору некоторой информации из различных источников
- в мобильных приложениях, где важно минимизировать траффик
- системах data mining(data mining — это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.)

Задача очистки веб-страниц от информационного шума весьма актуальна в наше время и решение данной проблемы поможет преподносить искомую пользователем информацию в удобном для него виде, а так же положительно скажется на результатах web-поиска, классификации информации, извлечение текстовой информации и т.п.

Анализ методов выделения основного веб-контента

Применяемые методы анализа структуры web-страниц можно разделить на:

1. Методы, основанные на выделении повторяющихся для всех (или части) страниц сайта фрагментов информации [1]
2. Методы, основанные на анализе dom-деревьев страниц сайта [3]
3. Комбинированные методы [2]
4. Методы синтаксического и визуального анализа[5]
5. Методы анализа страниц построенных на HTML 5

Анализ существующих методов выделения основного веб-контента показал, что методы, основанные на анализе DOM дерева эффективны и просты, а также предоставляют возможность проводить обработку единичной веб-станицы.

Существуют инструментальные средства, которые частично решают задачу выделения основного веб-контента: Adblock Plus, NoScript, FlashBlock, Safari Reader, Readability. Все эти средства, в основном направлены на борьбу с рекламой. Проведенный обзор существующих инструментальных средств очистки веб-страниц от информационного шума позволил выделить основные трудности, с которыми сталкиваются пользователи:

- Блокирование полезного для пользователя контента.
  Зачастую системы выделения основного контента вместе с навигацией и баннерами блокируют и полезную информацию для пользователя (например, ссылки на сопутствующие статьи и прочее), причем пользователю данная информация станет доступной лишь при отмене обработки веб-страницы
- Не универсальность
  Множество существующих средств разработаны под конкретный браузер, что приводит к сужению категории пользователей
- Отсутствие адаптации под конкретного пользователя
  Обзор показал, что при работе выделения основного контента веб-страницы инструментальные средства основываются на общем восприятии понятия «полезная информация» - блок текстовой информации, что не всегда соответствует запросам пользователя
- Недостаточная эффективность

Исходя из всего вышесказанного, можно сделать вывод, что разработка инструментальных средств очистки веб-страниц от информационного шума ведется довольно активно, но пока не существует универсальных средств, которые бы могли удовлетворить все запросы пользователей.

Постановка задачи разработки классификационной схемы

Анализ вопроса очистки web-страниц от информационного шума дал возможность определить несколько типов сайтов исходя из соответствующих им характерным признакам и их значения. Типы сайтов и признаки характерные для них были сведены в табл.1

Таблица 1 – Признаки типов сайтов

Типы сайтов	Характерные признаки
- Фотогалереи - Фотосайты - Интернет магазины - Видеосайты	Для этих сайтов характерно высокое количество информативных изображений, т.е изображений которые будут полезны для пользователя и могут считаться полезным контентом
- Торренты - Поисковые системы и модули - Интернет магазины	Для данных сайтов характерно высокое количество ссылок, которые могут считаться полезным контентом

Для разработки классификационной схемы необходимо определить структуру DOM-дерева HTML-страницы.

На сегодняшний момент выделяют три вида построения структуры веб-страницы:

1) Табличная (TABLE).
   Самая старая и распространенная структура сайтов.
2) Современная блочная (DIV)
   При создании сайта применяется современная блочная структура, которая имеет ряд преимуществ перед обычной HTML верстkой сайтов.
3) Структура с использованием FRAME
   Фреймы используются для разбивки окна браузера на несколько областей, каждая из которых представляет собой отдельный HTML-документ (фрейм). Как правило, фреймы используются для облегчения навигации по сайту, создания навигационного меню. Тем не менее, большинство разработчиков избегают использования фреймов, поэтому в дальнейшем этот вид структуры сайта рассматриваться не будет.

Стоит также отметить, что существуют web-страницы, структура которых комбинирует в себе использование TABLE и DIV.

Таким образом, будем считать, что очищенная страница от информационного шума представляется в виде:

$$S' = F(S, b), \qquad (1)$$

где $F(S, \bar{b})$ – функция очистки,

S - исходный сайт,

$\bar{b}$ - вектор, который определяется набором следующих признаков.

В пределах статьи будут рассматриваться следующие признаки вектора $\bar{b}$

KImg - количество изображений,

KObject -количество flash контента,

KLink - количество гиперссылок ,

KLists -количество таких тегов как <ul>,<ol>,<li>.

Для каждого признака необходимо определить интервал значений, который будет считаться нормой. Отклонения за максимальную границу интервала будем принимать как признак информационного шума, который требует очистки.

Разработка алгоритма очистки от информационного шума

Приняв во внимание все сильные и слабые стороны существующих инструментальных средств, остановим свой выбор на идеи создания букмарклета.

Букмарклет(bookmarklet) - это javascript-код, который сохраняется как закладка в браузере. Он работает за счет использования протокола <a href="javascript:...">.

Алгоритм очистки web-страниц от информационного шума состоит из следующих этапов:

1. Букмарклет получает адрес страницы для ее обработки.
2. Для заданной страницы определяется структура DOM дерева из HTML-кода.
3. Происходит проход по DOM дереву и классификация тегов(узлов) по соответствующим признакам.
4. Далее определяются значимые узлы.
5. Система обрабатывает информационные блоки, выделяет блок основного контента, отсекая теги, помеченные как информационный шум (медиа, навигация, ссылки и прочее).
6. Обработанная страница отображается для пользователя.
7. В случае, если произошло отсечение важной информации, пользователь отменяет обработку. Страница отображается ему в первичном виде с рамками вокруг различных блоков контента. Отметив нужный блок, пользователь сохраняет

результат. Страница вновь проходит обработку, в ходе которой отмеченные пользователем блоки отсекаться не будут. Обработанная страница отображается для пользователя вместе с сообщением, в котором будет предложено сохранить результаты обработки страницы в системе.

8. Адрес обрабатываемой страницы и результаты ее обработки сохраняются.

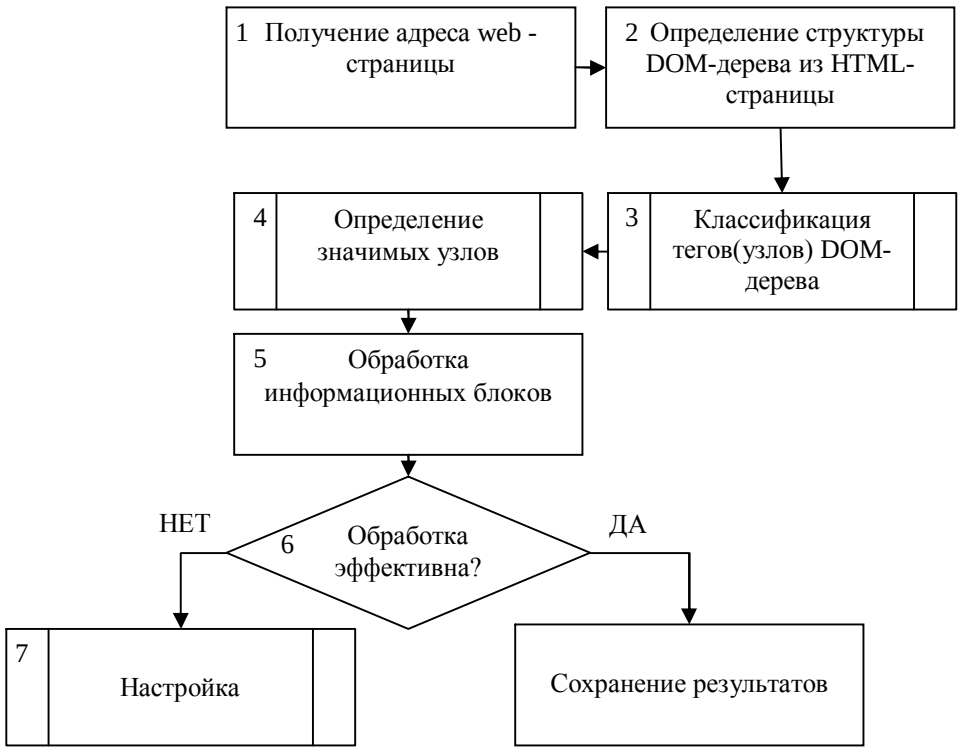

Рис. 1 – Блок-схема алгоритма работы букмарклета

Экспериментальное исследование определения значений параметров информационных блоков

Для выделения информационных блоков необходимо разработать средство позволяющее работать с html-кодом страницы непосредственно в окне браузера. Кроме этого для определения типа информационного блока необходимо рассчитать значения его параметров.

Для исследования был разработано специальное программное обеспечение - букмарклет, который выделяет div и table верхнего уровня, определяет количество заданных признаков в пределах структурного блока и по странице в целом. Набор исследуемых параметров определен в постановке задачи.

Исследование проводилось по следующей методике: для каждого из 10 поисковых запросов из разных областей было загружено по 20 первых веб-ресурсов, выданных поисковой системой Google. Каждая страница была проанализирована и разбита на структурные блоки (div и table верхнего уровня). Для каждого блока и для всей web-странице в целом было посчитано количество изображений, ссылок, списков и flash-объектов. В результате выше перечисленных действий было получена выборка, состоящая приблизительно из 500 записей. Пример использования букмарклета для проведения исследования показан на рис.2.

На основе вычисленных характеристик выделим границы каждого признака значения, внутри которых будет считаться нормой, а в случае отклонения от максимальной границы признак будет признан информационным шумом. Вычисленные значение приведем в табл.2.

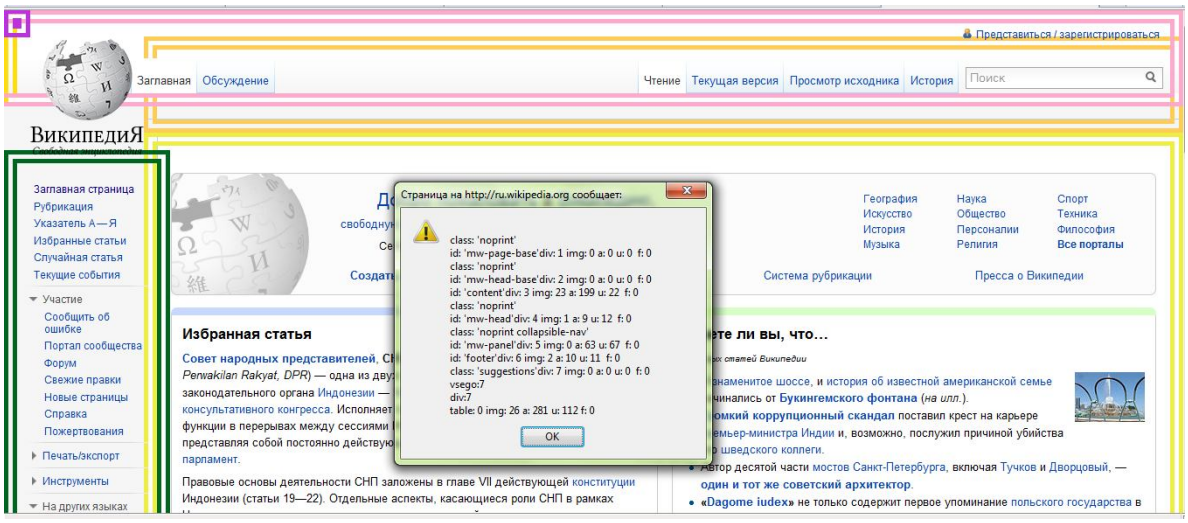

Рис. 2 – Пример использования букмарклета

Таблица 2 – Границы характерных признаков для различных типов сайтов

Тип сайта	Признак информационного блока					
	Количество изображений		Количество ссылок		Количество элементов списков	Количество flash-объектов
	min	max	min	max	max	max
Фотосайты	12	195	22	1089	120	1
Видеосайты	14	277	96	1177	336	4
Торренты	12	120	88	1885	216	3
Поисковые модули	1	71	27	270	213	1
Интернет магазины	22	296	72	1005	838	4
Обычные сайты	0	289	10	3218	320	2

Выводы

В статье рассмотрена проблема определения основного контента web-страницы, который будет полезен для пользователя. Предложены статические характеристики, в зависимости от которых будет определяться значимость информационных блоков, а так же определены специфические виды сайтов, для которых статические характеристики будут отличны от обычных web-страниц.

Литература

1. И. Некрестьянов, Е. Павлова. Обнаружение структурного подобия HTML-документов. СПГУ, 2002. - С. 38 – 54. – http://meta.math.spbu.ru

2. Р.Ф. Кузнецов, Н.В. Мурашов. Оценка влияния извлечения значимой информации на качество классификации web-страниц

3. Soumen Chakrabarti. Integrating the Document Object Model // In Proceedings of WWW10, May 1-5, 2001,[электронный ресурс]. Режим доступа: http://www10.org/cdrom/papers/489

4. Краковецкий А. Очищаем веб-страницы от информационного шума [электронный ресурс]. Режим доступа: http://msug.vn.ua/blogs/datamining/archive/2009/08/06/1010.aspx