УДК 519.237+004.4:234

РАЗРАБОТКА АДАПТАЦИОННОГО АЛГОРИТМА ОЧИСТКИ ВЕБ-СТРАНИЦ ОТ ИНФОРМАЦИОННОГО ШУМА

Криницкая А.И., Мартыненко Т.В.
Донецкий национальный технический университет
кафедра автоматизированных систем управления

В работе обсуждаются основные категории пользователей сети интернет, определена целевая аудитория для системы очистки веб-страниц от информационного шума. Разработан адаптационный алгоритм для системы очистки веб-страниц от информационного шума. Определены основные проблемы существующей системы. Приведен предлагаемый алгоритм 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» [4]. К нему можно отнести навигацию, связанные ссылки, элементы дизайна, рекламу. Весь этот «информационный шум» зачастую мешает нормальному восприятию необходимой информации.

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

Цель работы – создание общедоступных инструментальных средств, позволяющих очистить web страницы от информационного шума.

Для достижения поставленной цели необходимо решить основные задачи:

1. Провести сравнительный анализ целевой аудитории пользователей системы очистки веб-страниц от информационного шума.
2. Определить основные категории пользователей среды Интернет исходя из интересующей их информации.
3. Разработать адаптационный алгоритм для системы очистки веб-страниц от информационного шума.

Планируемая практическая значимость работы заключается в адаптация представления информации на сайте под запросы пользователя

Математическая постановка

Очищенная страница от информационного шума представляется в виде:

$$S' = F(S,b)$$

где $F(S,b,l)$ – функция очистки; S – исходный сайт; l – адаптивный признак пользователя, задается перечнем контрольных слов для каждой группы пользователей; b – параметр блока контента, который определяется по следующей формуле:

$$b = \begin{cases} 0, \text{если } g(K\,Im\,gs, KSents, KStopWords, KObject, \overline{p_0}) = TRUE; \\ 1, \text{если } g(KLinks, KText, KLists, KNav, \overline{p_1}) = TRUE; \\ 2, \text{если } g(KMedia, KSents, KTextSents, \overline{p_2}) = TRUE, \end{cases}$$

где g – функция определения свойств содержимого сайта; $\overline{p_i}$ – параметры обработки для различных информационных блоков

Анализ аудитории пользователей системы

Необходимость автоматического анализа информации из интернета вызвана высокой доступностью огромного количества постоянно пополняющейся информации, а также растущей популярностью веб-услуг среди всех категорий пользователей.

Система очистки веб-страниц от информационного шума ориентирована на возрастную категорию от 16 лет и старше. Предполагаемое возрастное ядро аудитории от 22 до 45 лет. Региональная принадлежность, пол, социальный статус для пользования этой системой не имеет. Требуется уровень образования достаточный для свободного использования ПК.

Учитывая то, что система рассчитана на огромное количество пользователей и может использоваться в разных сферах деятельности необходимо разработаь алгоритм адаптации системы под нужды конкретного пользователя.

Для достижения поставленной цели были решены следующие задачи:

- Категоризация потенциальных пользователей системы.
- Определены соотношения между категорией пользователя и интересующей информацией.
- Предложен алгоритм адаптации системы под пользователя.

Основные категории пользователей среды сети Интернет

Среда глобальной информационной сети Интернет, которую именуют «киберпространством», «социальной виртуальной реальностью», «нулевым» пространством, «параллельным» миром, новой «средой обитания», комплексным экологическим пространством, «информационным слепком человеческого бытия», «новым жизненным пространством», «второй реальностью», «субъективной реальностью» - это не только взаимосвязанные посредством коммуникационного оборудования, технологий и программ компьютеры, а, прежде всего взаимодействующие в этой среде люди [1].

В настоящее время существуют множество классификаций пользователей сети Интернет по различным основаниям. Приведем описание тех, которые наиболее отвечают тематике работы в табл. 1.

В табл. 1 приведены возможные значения адаптивного признака категории пользователя. Данные значение представляют собой набор слов, которые являются поисковым запросом пользователя. Именно набор значений адаптивного признака позволит отнести пользователя к определенной категории.

Таблица 1. Основные категории пользователей Интернет

l	Категория пользователей	Описание категории	Значения адаптивного признака категории
1	Блоггер	Для данной категории пользователей основным контентом будет считаться большое количество текстовой информации, ссылки на записи, умеренное количество изображений.	новости, отзывы, форум, чат, общение, пост, жж, интересно, как сделать, почему, ответ, история, блог
2	Потребитель	Для данной категории пользователей характерна активность на сайтах размещающих разнообразное программное обеспечение доступное для общего пользования.	скачать, скачать бесплатно, скачать фильм, скачать сериал, скачать книгу, скачать карты, скачать программу, скачать реферат
3	Покупатель	Категория пользователей, которые пользуется интернетом для покупок, оплаты счетов, заказов и онлайн бронирования. Важной информацией для данной категории можно считать умеренное количество изображений, ссылки на сопутствующие товары и т.п.	купить, активировать, забронировать, интернет магазин, онлайн бронирование
4	Бизнесмен	Категория пользователей, которые ищут полезную информацию, интересуются исследовательскими проектами, в поисковых запросах присутствуют слова «новости», «читать», «описание», «исследование» и прочее.	публикация, авторство, новости, исследование, открытие, обнаружить, проект, методичка, пособие
5	Обычный пользователь	Категория для пользователей, чьи интересы неоднозначны и не могут быть установлены. В данную категорию будут относится пользователи которые не соответствуют ни одной из выше перечисленных категорий.	-

Определение ключевых слов поисковых запросов пользователей

Для определения ключевых слов запросов пользователей используются данные cookie и HTTP referer.

Referer (HTTP referer) – в протоколе HTTP один из заголовков запроса клиента. Содержит URL источника запроса. Если перейти с одной страницы на другую, referer будет содержать адрес первой страницы. Часто на HTTP-сервере устанавливается программное обеспечение, анализирующее referer и извлекающее из него различную информацию. Так, например, владелец веб-сайта получает возможность узнать, по каким поисковым запросам, как часто и на какие именно страницы попадают люди [2].

Cookie – фрагмент данных, созданный веб-сервером или веб-страницей и хранимый на компьютере пользователя в виде файла, который веб-клиент (обычно веб-браузер) каждый раз пересылает веб-серверу в HTTP-запросе при попытке открыть страницу соответствующего сайта [3]. Применяется для сохранения данных на стороне пользователя, на практике обычно используется для:

- аутентификации пользователя;
- хранения персональных предпочтений и настроек пользователя;
- отслеживания состояния сессии доступа пользователя;
- ведения статистики о пользователях.

В рамках данной работы для нас представляют важность cookie _utmz. Благодаря им предоставляется возможность определить как пользователь оказался на сайте, если он воспользовался ссылкой с другого ресурса, можно установить с какого именно и по каким ключевым словам он искал сайт, если пришел с поисковика.

Срок жизни « 6 месяцев, обновляются при загрузке очередной страницы сайта.

`Формат: XXXX.TTTT.V.S.utmcsr{source}|utmccn{campaign}|utmcmd{medium}|utmctr{keyword}`

Значения:

- XXXX – hash домена.
- V – количество посещений пользователем сайта, совершенных по ссылкам с других ресурсов.
- S – количество различных ресурсов, с которых пользователь попадал на сайт.
- utmcsr – ресурс-поисковик, с которого пользователь попал на сайт.
- utmccn – содержит информацию о компании из AdWords (или значение utm_campaign в запросе) или же сообщает, что пользователь попал к вам посредством organic search.
- utmcmd – содержит название компании (или значение utm_medium в запросе) или сообщает об organic search.
- utmctr – ключевые слова, по которым велся поиск.

Адаптационный алгоритм очистки веб-страниц от информационного шума

Адаптация очистки веб-страницы под конкретного пользователя заключается в следующих этапах:

1. Пользователь активировал систему очистки веб-страницы от информационного шума.
2. С обрабатываемой страницы в систему передается информация, содержащаяся в cookie.
3. Система считывает полученную информацию, определяет ключевые слова по которым пользователь перешел на обрабатываемый сайт.
4. Система анализирует полученный набор слов, сопоставляет их с наборами ключевых слов определенных по каждой категории пользователей.
5. При обнаружении совпадений система относит пользователя к определенной категории пользователей. Стоит отметить, что сравниваются всего 4 набора ключевых слов с полученными данными, в то время, как категорий пользователей пять.
6. В случае когда пользователя нельзя отнести ни к одной категории обладающей набором

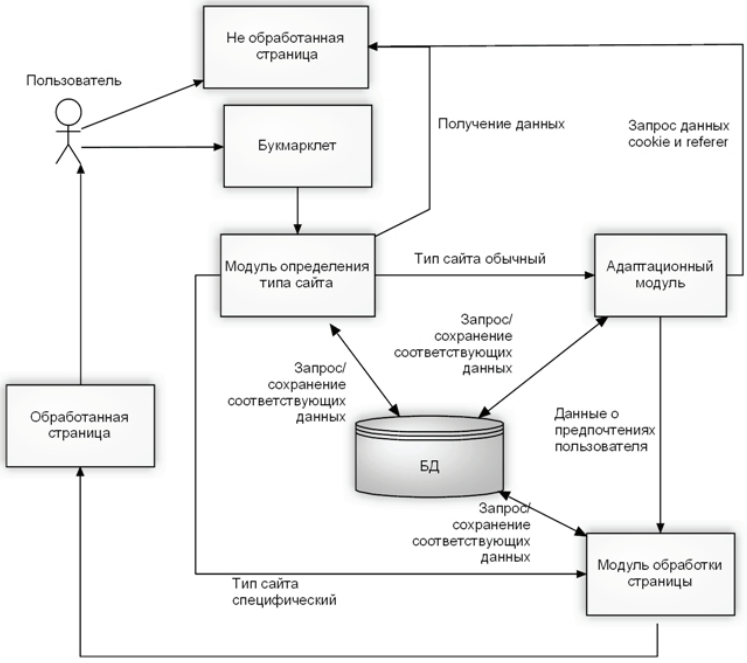

Рисунок 1 – Схема работы системы очистки веб-страниц от информационного шума

ключевых слов (потребитель, покупатель, бизнесмен, блоггер) система относит пользователя к категории «обычный пользователь».

7. Происходит обработка страницы исходя из заданных параметров выбранной категории.
8. Данные сохраняются в базу данных по сайту и по категории к которой относится пользователь.

Обобщенная схема работы системы очистки веб-страниц от информационного шума представлена на рис.1.

Выводы

В статье рассмотрена проблема адаптации системы очистки веб-страниц от информационного шума под конкретного пользователя. Разработан адаптационный алгоритм для системы выделения основного контента.

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] Определение понятия «информационный шум» [электронный ресурс]. Режим доступа: http://mediart.ru/blog/kiberzhurnalistika/742-1-pered-viborami-kak-upravlyat-smi-chtobi-oni-ne-upravlyali-vami.html