УДК 004.78:025.4.036

МЕТОД РАЗБИЕНИЯ ВЕБ-СТРАНИЦ НА СЕМАНТИЧЕСКИЕ БЛОКИ С ЦЕЛЬЮ ВЫЯВЛЕНИЯ СХОЖИХ ДОКУМЕНТОВ

Д. И. Косинов

Воронежский государственный университет

Поступила в редакцию 20.05.2008 г.

Аннотация. Задача поиска схожих документов рассмотрена с точки зрения составляющих их блоков. Предлагается алгоритм, позволяющий выделять семантические блоки из web-страниц путем анализа DOM-дерева. Предлагается метод, позволяющий поблочно определять похожесть web-документов между собой и дающий общую оценку степени схожести многоблочных документов. Показан прирост качества распознавания дубликатов на основе сравнения метода шинглирования.

Ключевые слова: web-страницы, семантические блоки, метод шинглирования, анализ DOM-дерева.

Abstract. In this paper we focus on selecting similar documents by analysing their constituent parts. An algorithm that selects semantic blocks from web-pages by analysing their DOM structure is proposed. A method that helps to determine the similarity of web-pages block-by-block and to estimate the degree of similarity of multiblock documents is proposed. It is shown that this method performs better than the shingles approach.

Key words: web-pages, semantic blocks, shingles approach, analysing DOM structure.

ВВЕДЕНИЕ

В связи с колоссальным объемом информации в Интернете и ее избыточностью из-за многочисленных страниц со почти одинаковым содержанием возникает одна из сложнейших задач анализа web-данных — проблема уверенного обнаружения схожих между собой web-документов. Существующие методы детектирования в большинстве своем основаны на генерации одного или нескольких отпечатков web-страницы, построенных с целью обеспечить максимальную устойчивость к небольшим изменениям содержимого. Дальнейший поиск совпадающих отпечатков в общей коллекции позволяет выделить документы, с большой долей вероятности являющиеся нечеткими дубликатами.

Традиционная процедура подготовки документа к созданию отпечатка заключается в конкатенации всего текстового содержимого и последующей обработке, невзирая на принадлежность текста к какому-либо логическому сегменту документа. Было высказано предположение, что предварительное разбиение текста на семантические блоки в соответствии со структурой документа позволит улучшить результативность стандартных методов создания отпечатков. В качестве аналогии можно привести работу [1], авторы которой добились некоторого улучшения результатов смежной задачи информационного поиска путем отделения навигационной части web-страницы от содержательной.

В данной работе предлагается метод, позволяющий поблочно определять похожесть web-документов между собой и дающий общую оценку степени схожести многоблочных документов.

1. ОПИСАНИЕ МЕТОДА

Рассматриваемый метод расширяет стандартные методы обнаружения схожих документов внесением в них понятия «блоков» страниц. В остальном последовательность действий остается стандартной: на первом этапе заполняется база отпечатков информационных блоков для каждого документа коллекции, на втором происходит непосредственное сравнение отпечатков и поиск схожих документов. Перечень этапов предлагаемого метода приведен на рис. 1.

Задача разбиения документа на семантические блоки сама по себе весьма неоднозначна и сложна. В Интернете доступно огромное количество информации, обладающей определенной

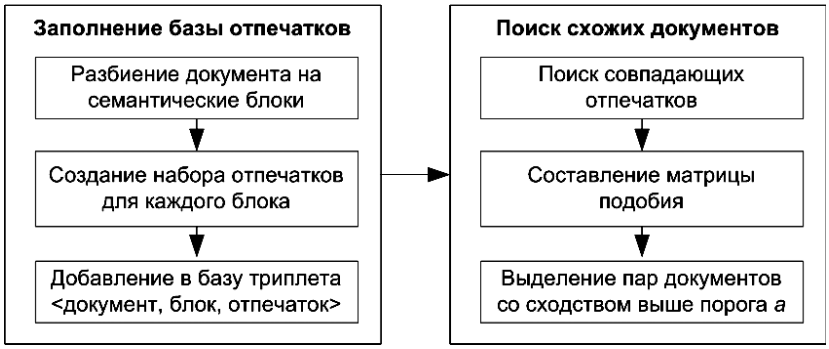

Рис. 1. Перечень этапов предлагаемого метода

структурой и возможность извлекать данные из таких структур, несомненно, полезна. Однако отсутствие априорных знаний о структуре хранения данных, необходимость фильтрации элементов оформления и прочие проблемы делают эту задачу достаточно сложной. Существуют различные подходы к извлечению информации из web-страниц, большинство из которых подробно рассматривается в работе [2].

В простейшем случае документы делятся на фрагменты одинакового объема. Это было приемлемым решением в условиях нехватки ресурсов для обработки больших документов, но на сегодняшний день только ухудшает результаты за счет возможного повреждения целостности важных областей.

Альтернативный подход использует информацию о частоте встречаемости искомых термов в различных фрагментах текста, что позволяет выделить наиболее значимые с точки зрения заданного набора блоки текста. Однако при отсутствии заранее заданного набора значимых термов этот подход неприменим.

Предлагаемый в [3] метод использования автоматических средств извлечения информации (так называемых «посредников», wrappers) подразумевает относительную регулярность набора документов, то есть наличие у них схожей структуры. Показывая высокое качество извлечения при работе с однотипной коллекцией, данный подход проигрывает в условиях «хаотичного» посещения web-страниц поисковым роботом.

Кроме того, существует ряд методов, основанных на выделении повторяющихся фрагментов web-страниц. Они исходят из предположения о том, что в пределах одного сайта страницы имеют повторяющиеся фрагменты в части навигационных блоков. В результате отбрасывания таких повторяющихся блоков остается значимая, содержательная часть документа. В данной работе эти методы не рассматриваются, так как требуют многократного прохода по коллекции документов для выделения содержательной части.

Последняя группа методов работает путем анализа структуры web-страницы [2, 3]. Стандарт DOM [4] формализует описание web-страниц в виде дерева html-тегов, что позволяет в автоматическом режиме строить DOM-дерево для заданного документа. Именно этот подход был применен в данной работе.

Блок-схема используемого алгоритма приведена на рис. 2.

Для каждого узла полученного после первого шага алгоритма дерева выполняется расчет веса. Вес узла зависит от объема контента (здесь и далее под контентом подразумевается текстовая информация, непосредственно выводимая

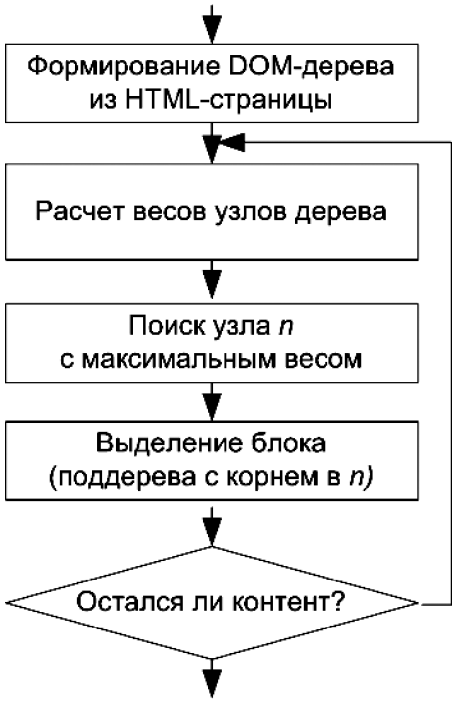

Рис. 2. Блок-схема алгоритма выделения блоков

на экран в результате отрисовки web-страницы браузером) его поддерева, а так же от динамики изменения объема контента при переходе на более высокий уровень. Исходя из этих соображений, весовая функция узла была определена следующим образом:

$$w = \frac{l \log(lk_1) \log(dk_2 + c)}{pk_3},$$

где l — суммарная длина контента поддерева, корнем которого является рассматриваемый узел, p — суммарная длина контента поддерева, корнем которого является родительский узел рассматриваемого узла, d — глубина узла (уровень вложенности), k_{1-3} и c — коэффициенты. Введение в функцию веса показателя глубины узла было продиктовано желанием выделять блоки по возможности на более низких уровнях.

Необходимо отметить, что введением дополнительных параметров узла (количество ссылок, число «структурных» тегов и т.п.) можно добиться выделения семантических блоков иной направленности: например, ссылочных (с высоким содержанием гиперссылок, свойственно для навигационных блоков) или высокоструктурированных (табличных).

После получения итогового текста блока производится создание его отпечатков. Среди множества алгоритмов создания отпечатков документов для использования в данной работе был выбран классический алгоритм шинглирования, описанный в [5]. Выбор был обусловлен ярко выраженной «неразборчивостью» данного алгоритма — последовательности слов, из которых получаются шинглы, с заметной долей вероятности берутся прямо на границе семантических блоков, соответственно включая в себя различные смысловые фрагменты. Шинглируя же блоки по отдельности, этой проблемы удастся избежать.

Каждый выделенный блок шинглировался в соответствии с заданными параметрами алгоритма шинглирования. Итоговый триплет следующего формата — [документ, блок, отпечаток] — добавляется в общую базу отпечатков.

После заполнения базы отпечатков для всех документов коллекции можно переходить ко второму этапу: поиску схожих документов. Сходство документов определяется через сходство отпечатков:

$$Sim(A,B) = \frac{|S(A) \cap S(B)|}{|S(A) \cup S(B)|},$$

где $S(A)$ — множество шинглов страницы A.

2. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТА

Построение DOM-дерева из web-страниц осуществлялось с помощью свободно распространяемой библиотеки HTMLParser[1] . Доля документов с некорректно распознанной html-разметки составляла не более 4 %, причиной чему в большинстве случаев служили логические ошибки в исходном тексте документа.

Для оценки качества выделения блоков использовалась вспомогательная коллекция из 20 web-страниц, подобранных из соображений максимальной разнообразности с целью охватить как можно большее количество различных вариантов структурирования контента.

Эксперименты с выделением блоков из документов на вспомогательной коллекции показали большую зависимость числа и содержания блоков от значений коэффициентов k_{1-3} и c. Наиболее корректных результатов удалось добиться, используя следующую комбинацию значений: $k_1 = 0.1$, $k_2 = 0.9$, $k_3 = 1$, $c = 1$. В этом случае на одну web-страницу приходилось 3—7 блоков, что является оптимальным диапазоном с точки зрения и наглядности, и количества получаемых на следующем этапе отпечатков. Число документов, разделенных на большее число блоков, не превышало 10 %. Пример работы алгоритма приведен на рис. 3.

Основной недостаток предложенного алгоритма выделения блоков вытекает из его формулировки: в результате того, что блоком считается поддерево с корнем в одном узле, возможно выделение слишком большого числа блоков в случае, если дерево является «плоским» (имеет небольшое число уровней глубины относительно большого числа однотипных узлов на одном уровне). Решением может послужить модификация алгоритма, позволяющая выделять в качестве блока сразу несколько поддеревьев. Эта задача, однако, является темой для отдельного исследования.

В качестве тестовой коллекции для проведения сравнительного анализа качества распознавания схожих документов использовалась специально созданная коллекция web-страниц сайта Воронежского государственного университета[2]. Она состояла из 160 файлов, доступных для скачивания с титульной страницы сайта в диапазоне не более, чем двух переходов по ссылкам.

[1] http://htmlparser.sourceforge.net/

[2] http://www.vsu.ru

Рис. 3. Результат работы алгоритма выделения блоков на титульной странице сайта www.vsu.ru

Было сделано два тестовых пробега: без разбиения на блоки и с разбиением. Параметры шинглирования задавались следующие: длина шингла — 10 слов, число шинглов на документ — 10. Число шинглов на блок было пропорционально отношению длины блока к общей длине документа, но не менее одного на блок.

В результате, несмотря на одинаковое число шинглов в обоих пробегах, число совпадений между шинглами документов во втором случае увеличилось более чем вдвое. Большая часть прироста произошла за счет навигационных частей страниц, уверенно выделяемых в процессе разбиения страниц на блоки. Таким образом, отдельное шинглирование этих идентичных блоков дало совпадения их шинглов более чем для 90 % страниц, что подтвердило первоначальные предположения.

ЗАКЛЮЧЕНИЕ

Результаты проведенного эксперимента показывают возможность принципиальной применимости описанного метода. Предложенный метод позволяет в том числе проводить сравнения не только документов в целом, но и на более детальном уровне составляющих их семантических блоков. К примеру, показана высокая эффективность в плане выделения навигационной части страниц, поскольку подобные навигационные блоки практически всегда выделяются разработчиками страниц в отдельную структурную единицу DOM-дерева.

В ходе дальнейших исследований планируется модернизировать алгоритм для корректной работы с «плоскими» деревьями, а так же определить оптимальные значения коэффициентов формулы веса для выделения блоков оптимального размера.

СПИСОК ЛИТЕРАТУРЫ

1. *Агеев М.С., Вершинников И.С., Добров Б.В.* Извлечение значимой информации из web-страниц для задач информационного поиска. “Интернет-Математика-2005”: семинар в рамках Всеросс. науч. конф. RCDL’2005 — Яндекс, 2005. — С. 283—301.

2. *Sephorah G.* Automatic Extraction of Generic Web Page Components. — (http://stp.ling.uu.se/exarb/arch/2004_graves.pdf).

3. *Некрестьянов И., Павлова Е.* Обнаружение структурного подобия HTML-документов. Труды четвертой всероссийской конференция RCDL’2002, Т. 2, С. 38—54, Дубна, Россия, 2002.

4. *Document Object Model (DOM) Level 2 HTML Specification*. — (http://www.w3.org/TR/DOM-Level-2-HTML/).

5. *Broder A., Glassman S., Manasse M., Zweig G.* Syntactic clustering of the Web. Proc. of the 6th International World Wide Web Conference, April 1997.

Косинов Дмитрий Иванович — ассистент кафедры программирования и информационных технологий факультета компьютерных наук Воронежского государственного университета. Тел. (4732) 208-470, E-mail: kosinov@cs.vsu.ru.

Kosinov D. I. — Post-graduate student, of the dept. of Programming and Information Technologies, Voronezh State University. Tel. (4732) 208-470, E-mail: kosinov@cs.vsu.ru.