Извлечение значимой информации из web-страниц с использованием предложений

© Р.Ф. Кузнецов

Балтийский Государственный
Технический Университет
ruslkuznetsov@gmail.com

Аннотация

Целью данной работы является разработка метода позволяющего отделять значимую часть web-страницы от навигационной, в качестве эвристики использую законченные предложения.

1. Введение

Навигационная часть web-страницы состоит из множества элементов, таких как ссылки на другие страницы сайта, блоки текста с рекламой, контактные данные компании и прочая служебная информация. Часто эти элементы не имеют прямого отношения к теме страницы, поэтому их индексация может ухудшить качество информационного поиска [3, 4, 5].

В данной статье мы исследуем эффективность метода, позволяющего отделять значимую часть web-страницы от навигационной. Особенностью метода является возможность выделения значимой части без использования информации с других страниц сайта.

2. Обзор существующих методов

В последнее время было опубликовано множество работ на тему выделения значимой части web-страниц. Описанные в них методы можно разбить на три основные группы:

1. методы, основанные на выделении повторяющихся для всех (или части) страниц сайта фрагментов информации [1, 2].
2. методы, основанные на анализе dom-деревьев страниц сайта [3, 5].
3. Методы, совмещающие оба этих подхода [4].

Методы, основанные на выделении повторяющихся фрагментов страниц одного сайта, являются более эффективными и универсальными. Однако для их работы необходима информация обо всех страницах сайта (или хотя бы части из них), что требует больших временных затрат и не всегда целесообразно.

3. Описание метода

В отличие от упомянутых выше подходов, рассматриваемый метод позволяет выделять значимую часть web-страницы без построения dom-дерева и анализа других страниц сайта.

Метод основан на двух предположениях: содержательная часть страницы обычно включает законченные предложения и вся информация, находящаяся в структурном блоке, содержащем предложения, относится к содержательной части. Структурным блоком будем называть фрагмент текстового содержания страницы, обрамленный тегами, не позволяющими сохранить единство предложения (например, <p>,
, <table> и т.п.).

Рассмотрим эти предположения подробнее.

В результате анализа HTML-страниц мы пришли к выводу, что в большинстве случаев наличие законченных предложений, является надежным признаком того, что текстовая информация относится к содержательной части страницы.

Далее, рассматривая структурные блоки, мы видим, что они относятся либо к навигационной, либо к содержательной части документа (то есть не принадлежат обеим частям сразу). Это происходит потому, что структурный блок также выполняет функции логического – делит текст страницы на структурно-однородные фрагменты.

Однако отсутствие в блоке предложения не гарантирует того, что он относится к навигационной части. Например, как правило, в конце заголовка web-страницы знаки препинания не ставятся. В этом случае мы можем использовать следующую эвристику: т.к. обычно заголовок страницы выделен относительного нижерасположенного текста (например, тегами <b>, <h1>, <h2>, <h3> и т.п. и/или размером шрифта) и не содержит ссылок, мы можем использовать эту информацию для поиска заголовков.

Алгоритм. При рассмотрении алгоритма можно выделить следующие этапы: представление HTML-документа, как совокупность структурных блоков; поиск предложений в каждом из блоков; поиск заголовков для блоков, отнесенных к содержательным.

Алгоритм разбивает HTML-документ на структурные блоки, используя следующие тэги - <p>,
, <td>, , <hr>, <form>. Далее в каждом блоке ищется хотя бы одно предложение. Поиск предложений основан на знаках препинания, таких как точка, восклицательный и вопросительный знаки. Алгоритм ищет точку после слова. «Слово» перед точкой может заканчиваться как буквой, так и одним из знаков препинания (например, ( ) {} [] ‘ ’ “ ” < >) или цифрой.

После того, как с помощью первых двух этапов работы алгоритма web-страница будет разделена на навигационную и содержательную части, происходит поиск содержательной информации среди блоков, не включающих предложения. Для блоков, приписанных к содержательной части, перед которыми расположены блоки из навигационной части, алгоритм ищет заголовки используя описанную выше эвристику.

4. Оценка эффективности работы алгоритма

Оценка эффективности алгоритма проводилась с помощью экспертного анализа. В качестве источника данных (web-страниц) была использована *«Веб коллекция Narod.Ru»*[1] содержащая более 22 000 сайтов и 728 000 страниц. Эксперт проанализировал результаты работы алгоритма для 113 случайно отобранных страниц.

Для конкретной страницы, в которой алгоритм выделял навигационную и содержательную части, эксперт выставлял одну из следующих оценок: «N-S-» - навигационная и содержательная части определены неправильно и включают части друг-друга; «N+S-» - навигационная часть включает только навигационную информацию; содержательная часть включает долю навигационной информации; «N-S+» - содержательная часть включает только содержательную информацию; навигационная часть включает долю содержательной информации; «N+S+» - навигационная и содержательная части определены правильно.

Результаты представлены в таблице 2.

Оценка	Количество web-страниц	Количество web-страниц %
N-S-	31	28
N+S-	14	12
N-S+	45	40
N+S+	23	20

Таб. 2. Результаты работы алгоритма

Рассмотрим ошибки алгоритма. Чаще всего они возникли при обработке таблиц, относящихся к содержательной части web-страницы; адресов; слоганов; нестандартно отформатированных предложений (разделенных между несколькими блоками); пунктов меню, состоящих из предложений или заканчивающихся точками и т.п. Таким образом, алгоритм наиболее часто ошибался тогда, когда принимал слова с точкой, как законченные предложения (например, адреса фирм) и когда содержательная часть страницы не включала предложения (например, ячейки таблиц).

Многих промахов в работе алгоритма можно избежать. Для этого необходимо разработать эвристики, с помощью которых можно корректно обрабатывать возникшие проблемные случаи.

Не смотря на существенную долю серьезных ошибок (28%) и ошибок, которые можно отнести к менее значимым (52%), анализируемый метод показал свою работоспособность и может рассматриваться как перспективный.

Заключение

Не смотря на вышеперечисленные недостатки, результаты эксперимента позволяют сделать вывод о применимости представленного алгоритма к выделению значимой части web-страниц и возможности повышения его эффективности.

Литература

[1] М.С. Агеев, И.В. Вершинников, Б.В. Добров. Извлечение значимой информации из web-страниц для задач информационного поиска. Интернет-математика 2005. Сборник работ по программам научных стипендий Яндекса. Москва, 2005.

[2] И. Некрестьянов, Е. Павлова. Обнаружение структурного подобия HTML-документов. Труды четвертой всероссийской конференции RCDL’2002, 38-54, Дубна, Россия, 2002.

[3] C. H. Lee, M.Y. Kan, S. Lai. Stylistic and Lexical Co-training for Web Block Classification. In Proceedings of Workshop on Web Information and Data Management (WIDM '04), Washington, D.C., USA.

[4] Yi, L., Liu, B., Web Page Cleaning for Web Mining through Feature Weighting, in the proceedings of Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03), Acapulco, Mexico, August, 2003.

[5] L. K. Shih and D. Karger. Using URLs and table layout for web classification tasks. In Proceedings of the 13th International Conference on the World Wide Web, pages 193--202, New York, NY, 2004.

Significant Part Extraction of Web Pages Using of Sentences

Ruslan F. Kuznetsov

The purpose of the given work is to check a hypothesis that in general finished sentences could be found, in a substantial part of a site. Using this hypothesis we have developed algorithm which divides web-pages into navigating and substantial parts.

[1] http://romip.narod.ru/ru/collections/narod.html