РАЗРАБОТКА ВЕБ-ПРИЛОЖЕНИЯ ДЛЯ СЕГМЕНТАЦИИ СОДЕРЖИМОГО HTML-СТРАНИЦ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

Алексеева А.А.[1]
[1)] Институт Вычислительной Математики и Информационных Технологий, Казанский (Приволжский) Федеральный Университет, Казань, Россия
E-mail: alexandrin-a@yandex.ru

DEVELOPMENT OF A WEB APPLICATION FOR SEGMENTATION OF HTML PAGE CONTENT USING MACHINE LEARNING METHODS

Alexeeva A.A.[1]
[1)] Institute of Computational Mathematics and Information Technologies, Kazan (Volga Region) Federal University, Kazan, Russia

This paper highlights the analysis of the main approaches to solving the problem of segmentation of an HTML page, and also provides a description of the algorithm for segmenting an HTML page based on semantic and structural analysis.

Часто люди, которые пишут научные работы, а также те, кто просто ищет информацию в интернете для себя, тратят большое количество времени на удаление лишней информации из скопированного текста. Проблема состоит в том, что на HTML-страницах большое количество рекламы и лишней информации, такой как комментарии, ссылки на другие страницы и т.д. Такая проблема может возникнуть не только при ручном изъятии текста, при автоматической обработке веб-страниц зачастую не лучшем решение является извлечение всей информации со страницы.

После анализа существующих статей по теме сегментации содержимого HTML-страниц были выделены три основных направления решения этой задачи: на основе разработанных правил, на основе анализа структуры элементов HTML-страницы, на основе семантического анализа HTML-страницы.

В решении задачи сегментации содержимого HTML-страниц важную роль играют и структура элементов HTML-страницы, и их содержимое. В связи с чем более перспективными являются решения, использующие оба подхода. Также задачу сегментации содержимого HTML-страниц можно свести к задаче классификации методами машинного обучения, как, например, описано в статье [1]. Авторы этой статьи с помощью структурного анализа DOM-дерева и семантического анализа выделяют набор признаков для каждого элемента HTML-страницы, и на основе полученного набора признаков для каждого элемента выноситься вердикт о принадлежности к классу «основной информации».

Цель данной работы заключается в том, чтобы разработать приложение, позволяющее выделить основную текстовую информацию из HTML-страниц. Для этого вся информация с HTML-страниц делится на три класса: «удаляемая информация», «основная информация» и «игнорируемая информация». Для решения задачи сегментации были реализованы два подхода: на основе структурного анализа DOM-дерева HTML-страницы и на основе одновременно и структурного, и семантического анализа. Оба алгоритма последовательно относят каждый элемент HTML-страницы к одному из трех классов. Для сравнения точности работы алгоритмов подсчитывались две оценки: среднее значение метрики accuracy разметки всех элементов HTML-страницы на три группы («удаляемая информация», «основная информация» и «игнорируемая информация») для 50 HTML-страниц из размеченного вручную набора данных и среднее значение метрики accuracy разметки всех элементов HTML-страницы на две группы («основная информация» и «не основная информация») для 100 HTML-страницах из преобразованного набора данных конкурса CleanEval. Алгоритм на основе одновременно и структурного, и семантического анализа показал улучшение обоих оценок более, чем на 10 %, по сравнению с алгоритмом на основе только структурного анализа.

1. Thijs Vogels, Octavian-Eugen Ganea, and Carsten Eickhoff. 2018. Web2text: Deep structured boilerplate removal. In European Conference on Information Retrieval, pages 167–179. Springer.