Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
Информатики и радиоэлектроники
Кафедра инженерной психологии и эргономики

На правах рукописи

УДК 004.774.6

Милюткин
Иван Николаевич

АЛГОРИТМ ОПРЕДЕЛЕНИЯ СЕМАНТИЧЕСКИХ БЛОКОВ ВЕБ-СТРАНИЦЫ

АВТОРЕФЕРАТ
Диссертации на соискание степени магистра технических наук

по специальности 1-23 80 08 Психология труда, инженерная психология, эргономика

Заведующий кафедрой ИПиЭ
Константин Дмитриевич Яшин
кандидат технических наук, доцент

Научный руководитель
Тамара Владимировна Казак
доктор психологических наук

Нормоконтролер
Татьяна Валерьевна Гордейчук
ассистент, магистр технических наук

Минск 2015

ВВЕДЕНИЕ

По мере роста и развития информационных технологий, а с ними и интернет-технологий, растёт и информационная нагрузка на пользователей. Всё чаще на различных сайтах увеличивается количество неинформативного контента, который можно назвать информационным шумом. Он не несёт никакой важной информации для пользователя.

Со временем это стало приводить к снижению эффективности восприятия информации, также ухудшило возможность сбора различной информации для систем анализа данных. Со временем аудитория, использующая компьютеры и Интернет, изменилась. В настоящее время пользователи хотят иметь насыщенные, но при этом простые в использовании пользовательские интерфейсы и чтобы информацию можно было легко получать необходимую информацию.

Объектом исследования являются семантические блоки веб-страниц. Предметом исследования является алгоритм определения семантических блоков веб-страниц.

Целью работы является разработка алгоритма для определения семантических блоков на веб-странице и его реализация. Основной задачей алгоритма является обеспечение точного и безошибочного определения семантических блоков на странице, а также их выделение.

Разрабатываемый алгоритм позволит автоматизировать процесс выделения фрагментов, предоставит механизм, который с одной стороны сможет определить фрагмент с разными темами или функциями, без участия человека, а с другой будет производительным и эффективным для обнаружения и маркирования таких же фрагментов на страницах с других веб-сайтов.

Для достижения поставленных целей ставим следующие задачи на исследование:

1 Провести подробный анализ области и существующих решений и алгоритмов по выделению семантических блоков на веб странице.

2 Разработать алгоритм выделения семантических блоков на веб-странице, который будет производительным и эффективным.

3 Реализовать алгоритм и приложение использующее алгоритм, описать возможности и места его применения.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Объектом исследования являются семантические блоки веб-страниц. Предметом исследования является алгоритм определения семантических блоков веб-страниц.

Целью работы является разработка алгоритма для определения семантических блоков на веб-странице и его реализация. Основной задачей алгоритма является обеспечение точного и безошибочного определения семантических блоков на странице, а также их выделение.

В первой главе магистерской работы производится анализ предметной области семантического веба и алгоритмов определения семантических блоков. Рассматриваются сферы применения алгоритма определения семантических блоков веб-страницы. Показываются особенности других алгоритмов. На основании проведённого анализа ставятся задачи на исследование.

Во второй главе проводится теоретическое исследование возможности разработки алгоритма определения семантических блоков, разработка и описание алгоритма. Обоснован выбор подхода к определению блоков.

В третьей главе производится техническая реализация алгоритма. Показаны примеры применения алгоритма, описаны возможности и области применения разработки.

ЗАДАЧИ

Задачей данного проекта является разработка алгоритма определения семантических блоков веб-сайта. Этот алгоритм должен уметь определять семантические блоки расположенные на веб-странице, а также выделять их. Данный алгоритм не должен зависеть от платформы, а также не должен быть требовательным к качеству разметки веб-страницы. Основными требованиями к алгоритмы являются:

– отсутствие зависимости от платформы;

– независимость от качества разметки;

– высокая производительность;

– возможность встраивать алгоритм в различные клиентские и серверные приложения;

– точность определения семантических блоков.

Для выполнения поставленных задач потребуется разработать алгоритм определения семантических блоков, а также выполнить его техническую реализацию.

ЗАКЛЮЧЕНИЕ

Разработанный в ходе выполнения исследования алгоритм определения семантических блоков позволяет разбивать страницу на различные блоки.

В результате проведения теоретического исследования и разработки алгоритма, в полном объёме описаны и получены базовые принципы, которые ложатся в основу разработки алгоритма.

Рассмотрена проблема определения основного информативного содержания веб-страницы. Определение блоков включает в себя разбиение веб-страницы на когерентные части и имеющие определённую функцию.

Рассмотрены и описаны особенности DOM-дерева и обоснование его необходимости для использования в алгоритме определения семантических блоков.

Разработана последовательность обработки веб-страницы алгоритмом, что позволит изолировать алгоритм от зависимости от платформы, а также описан путь применения алгоритма.

В результате реализации алгоритма определения семантических блоков, было реализована серверная и клиентская реализация разработанного алгоритма. Это свидетельствует о том, что алгоритм является платформонезасимым и реализуем с использованием основных современных технологий.

Определены дальнейшие пути развития разработанного алгоритма. Необходимо реализовать механизм определения назначения блоков. Это возможно реализовать с помощью нейронной сети.