БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Факультет радиофизики и компьютерных технологий

Кафедра интеллектуальных систем

Аннотация к магистерской диссертации

Применение машинного обучения для извлечения данных из веб-страниц

специальность 1-31 80 07 «Радиофизика»

Брагинец Илья Андреевич

Научный руководитель: Безродный Алексей Анатольевич, доктор технических наук, профессор

Минск, 2021

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Магистерская диссертация: 56 страниц, 15 рисунков, 22 таблицы, 50 источников.

СБОР ДАННЫХ, ВЕБ-СКРЭПИНГ, АВТОМАТИЗАЦИЯ, DOM, МАШИННОЕ ОБУЧЕНИЕ, КЛАССИФИКАЦИЯ.

Объект исследования – веб-страницы.

Предмет исследования – нахождение семантических элементов на вебстранице.

Цель работы – разработка метода построения автоматических (или автоматизированных, но с минимально возможным участием человека) систем сбора структурированных данных с веб-страниц, не требующих предварительной настройки для каждого веб-сайта.

Задачи:

• Рассмотреть методы анализа содержимого веб-страниц и выбрать наилучший по критерию минимальности ошибки автоматизированного распознавания.

• Собрать и подготовить данные для проведение вычислительных экспериментов.

• Показать возможность построения автоматической системы сбора структурированных данных на основе предложенного метода.

Методы исследования – DOM дерево, модели классификации машинного обучения.

В работе предложен метод, который позволяет извлекать данные с веб-сайта без предварительной настройки. Метод основан на анализе DOM (Document Object Model – объектная модель документа) дерева веб-страницы с помощью моделей машинного обучения.

Результаты экспериментов с веб-страницами интернет-магазинов показали, что у метода есть потенциал для построения универсальной системы сбора данных, которая позволяет извлекать данные с веб-страниц без предварительной настройки.

АГУЛЬНАЯ ХАРАКТАРЫСТЫКА РАБОТЫ

Магістарская дысертацыя: 56 старонак, 15 малюнкаў, 22 табліцы, 50 крыніц.

ЗБОР ДАНЫХ, ВЭБ-СКРЭПИНГ, АЎТАМАТЫЗАЦЫЯ, DOM, МАШЫННАГА НАВУЧАННЯ, КЛАСІФІКАЦЫЯ.

Аб'ект даследавання – вэб-старонкі.

Предмет исследования – знаходжанне семантычных элементаў на вэбстаронцы.

Мэта работы – распрацоўка метаду пабудовы аўтаматычных (або аўтаматызаваных, але з мінімальна магчымым удзелам чалавека) сістэм збору структураваных дадзеных з вэб-старонак, якія не патрабуюць папярэдняй налады для кожнага вэб-сайта.

Задачы:

- Разгледзіць метады аналізу змесціва вэб-старонак і выбраць найлепшы па крытэры мінімальна памылкі аўтаматызаванага распазнавання.
- Сабраць і падрыхтаваць дадзеныя для правядзенне вылічальных эксперыментаў.
- Паказаць магчымасць пабудовы аўтаматычнай сістэмы збору структураваных дадзеных на аснове прапанаванага метаду.

Метады даследавання – DOM дрэва, мадэлі класіфікацыі машыннага навучання. У працы прапанаваны метад, які дазваляе здабываць дадзеныя з вэб-сайта без папярэдняй налады. Метад заснаваны на аналізе DOM (Document Object Model – аб'ектная мадэль дакумента) дрэва вэб-старонкі з дапамогай мадэляў машыннага навучання.

Вынікі эксперыментаў з вэб-старонкамі інтэрнэт-крам паказалі, што ў метада ёсць патэнцыял для пабудовы універсальнай сістэмы збору дадзеных, якая дазваляе здабываць дадзеныя з вэб-старонак без папярэдняй налады.

GENERAL CHARACTERISTIC OF WORK

Master’s thesis: 56 pages, 15 figures, 22 tables, 50 sources.

DATA COLLECTION, WEB WRAPPERS, AUTOMATION, DOM, MACHINE LEARNING, CLASSIFICATION.

Object of research – web pages.

Subject of research – finding semantic elements on a web page.

Objective - development of a method for constructing automatic (or automated, but with the minimum possible human participation) systems for collecting structured data from web pages that do not require preliminary configuration for each website.

Tasks:

• Consider methods for analyzing the content of web pages and choose the best one based on the criterion of minimizing automated recognition errors.

• Collect and prepare data for computational experiments.

• Show the possibility of building an automatic system for collecting structured

data based on the proposed method.

Research methods - DOM tree, machine learning classification models.

The paper proposes a method that allows you to retrieve data from a website without prior configuration. The method is based on the analysis of the DOM (Document Object Model) of a web page tree using machine learning models. Experiments with e-commerce web pages have shown that the method has the potential to build a universal data collection system that allows you to retrieve data from web pages without prior configuration.