БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Факультет радиофизики и компьютерных технологий
Кафедра интеллектуальных систем

Аннотация к дипломной работе

«Приложение для агрегирования ресурсов сети Интернет»

Урбан Павел Владимирович

Научный руководитель: кандидат физико-математических наук, доцент
К. В. Козадаев

2017

РЕФЕРАТ

Дипломная работа: 79 страниц, 16 рисунков, 11 источников, 1 приложение.

ПОИСКОВЫЙ РОБОТ, АГРЕГИРОВАНИЕ ДАННЫХ, ИЗВЛЕЧЕНИЕ СОДЕРЖИМОГО ВЕБ-СТРАНИЦ, ПЛАНИРОВАНИЕ ОЧЕРЕДНОСТИ ЗАГРУЗОК ПОИСКОВОГО РОБОТА, ВРЕМЕННЫЕ ХАРАКТЕРИСТИКИ ВЕБ-СТРАНИЦ

Объект исследования – поисковый робот для агрегирования и анализа содержимого конечного списка веб-сайтов.

Цель работы – разработка поискового робота для агрегирования и помещения в базу данных содержимого веб-страниц для последующего анализа и вывода временной статистики их содержимого.

В результате выполнения работы изучены основы разработки поисковых роботов и возможности их применения, типичная архитектура поисковых роботов, рассмотрены сложности, возникающие при разработке поисковых роботов, такие как извлечение URL-адресов, планирование очередности сканирования, захвата обновлений. Разработано приложение на языке программирования Java для агрегирования и последующего анализа содержимого веб-страниц ограниченного списка веб-сайтов с авто объявлениями.

РЭФЕРАТ

Дыпломная праца: 79 старонак, 16 малюнкаў, 11 крыніц, 1 прыкладанне.

ПОШУКАВЫ РОБАТ, АГРЭГАВАННЯ ДАДЗЕНЫХ, ВЫМАННЕ ЗМЕСЦІВА ВЭБ-СТАРОНАК, ПЛАНАВАННЕ ЧАРГОВАСЦІ ЗАГРУЗАК ПОШУКАВАГА РОБАТА, ЧАСАВЫЯ ХАРАКТАРЫСТЫКІ ВЭБ-СТАРОНАК

Аб'ект даследавання – пошукавы робат для агрэгавання і аналізу змесціва канчатковага спісу вэб-сайтаў.

Мэта працы – ачыстка маўленчых сігналаў ад выпадковых шумоў як традыцыйнымі (з выкарыстоўваннем фільтраў), так і з выкарыстоўваннем вэйвлет-пераўтварэння і наступны аналіз атрыманых вынікаў.

У выніку выканання работы вывучаны асновы распрацоўкі пошукавых робатаў і магчымасці іх прымянення, тыповая архітэктура пошукавых робатаў, разгледжаны складанасці, якія ўзнікаюць пры распрацоўцы пошукавых робатаў такія як выманне URL-адрасоў, планаванне чарговасці сканавання, захопу абнаўленняў. Распрацавана прыкладанне на мове праграмавання Java для агрэгавання і наступнага аналізу змесціва вэб-старонак абмежаванага спісу вэб-сайтаў с аўто аб'явамі.

ABSTRACT

Thesis: 79 pages, 16 figures, 11 sources, 1 application.

WEB CRAWLER, DATA AGGREGATION, CONTENT EXTRACTION OF WEB PAGES, CRAWL ORDERING, TIME CHARACTERISTICS OF WEB PAGES

Object of research – web crawler for aggregating and analyzing the content of a limited list of websites.

Objective – development of a web crawler for aggregating and placing the contents of web pages in the database for later analysis and displaying the time statistics of their contents.

The basics of the development of web crawler and the possibilities of their application, the typical architecture of web crawler studied. Examined the difficulties arising in the development of web crawler such as extracting URLs, crawl ordering, capturing updates. An application has been developed in the Java programming language to aggregate and then analyze the contents of web pages of a restricted list of websites with auto ads.