УДК 004.91

РАЗРАБОТКА СИСТЕМЫ СБОРА ДАННЫХ С НЕСТРУКТУРИРОВАННЫХ СТРАНИЦ НА ПРИМЕРЕ САЙТА PEREKRESTOK.RU

Автор - Иванова Т.Н.

Научный руководитель — к.т.н., Штенников Д.Г.

федеральное государственное автономное образовательное учреждение
высшего образования
САНКТ-ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ,
г. Санкт-Петербург

Организациям необходимо использовать структурированные данные, собранные с ресурсов в сети Интернет, для дальнейшего применения в своих проектах. Такие данные не являются структурированными и требуют предварительной обработки, кроме того, информационные ресурсы которых осуществятся сбор данных могут быть неоднородными по своей природе (структура страниц, ссылок, навигационных элементов). Следовательно, получение структурированной открытой информации с веб ресурсов со слабой структурой страниц является сложной и актуальной задачей. Невозможность однозначно определить расположение и значение информации на странице может приводить к потере важных данных при извлечении. Дальнейшее использование неполных данных в других проектах может привести к проблемам в их работе.

Целью данной работы является разработка автоматизированной системы сбора данных с неструктурированных страниц. Разрабатываемая система позволит минимизировать потери важных данных и предоставить информацию в структурированном виде.

Базовые положения исследования. В работе рассмотрены ручные, гибридные и автоматические методы извлечения данных с веб ресурсов, обработка и структурирование полученных данных и хранение их в базе данных. Использованы различные инструменты автоматического извлечения данных для оптимизации процесса. Система также позволяет осуществлять обучение специалистов по датамайнингу на основе получения эталонных данных с веб-ресурсов и проверкой совпадения эталонной выборки с выборкой, выполненной в процессе обучения. Для оценки используются нечеткие множества.

Выводы. В результате проделанной работы была создана автоматизированная система сбора данных с неструктурированных страниц, которая минимизирует потери нужной информации и уменьшает время на их извлечение.

Автор: Иванова Т.Н. _____

Научный руководитель: Штенников Д.Г. _____

Руководитель образовательной программы: Горлушкина Н.Н._____