Т. С. Дубовик Е. М. Березовская
(ГГУ им. Ф. Скорины, Гомель)

АВТОМАТИЗАЦИЯ СБОРА ДАННЫХ С ПОРТАЛА TUT.BY

Все данные, представленные в глобальной сети Интернет, можно назвать неструктурированными, ввиду индивидуальности и специфичности архитектуры каждого ресурса. В основном, такие данные – это HTML страницы, т.е. текстовые структуры. В настоящее время, в связи с постоянным ростом информации во всемирной паутине, необходимо развитие технологий, позволяющих использовать ее для решения различных производственных задач предприятий и организаций, вследствие чего активно развивается область анализа текстовых данных и неструктурированной информации.

Существует множество систем для выполнения задач извлечения текстовых данных из интернет-источников [1]. Вот некоторые из них: TSIMMIS, WebOQL, FlORID, XWRAP, RoadRunner, Lixto, RAPIER, SRV, WHISK и др.

В предлагаемой работе используется технология Selenium. Selenium – это инструмент для автоматизированного управления веб-браузерами, однако он также отлично подходит для сбора информации с веб-страницы. Selenium представляет набор инструментов для работы с элементами веб-страниц, позволяет производить поиск элемента на странице с помощью CSS-описания, XPath-пути, имени класса элемента и др. [2,3]

Задачей данной работы является автоматизация сбора информации о курсах валют с белорусского информационного портала tut.by. Был разработан программный продукт, позволяющий практически моментально, получить свежую информацию о курсах валют. Для разработки была использована среда Microsoft Visual Studio 2013, для работы с элементами веб-страницы использовались технологии Selenium.WebDriver и XPath.

Разработанный продукт является лишь простейшим примером автоматизации сбора информации с веб-сайтов. Данное направление информационных технологий сейчас находится в непрерывном развитии, технологии автоматизации начинают использоваться в бизнес-целях, например, для сбора информации о ценах определённого продукта на различных сайтах и, следовательно, последующем ценообразовании для данного продукта.

Литература

1 Laender, A. H. F. A brief survey of web data extraction tools / A. H. F. Laender, B. A. Ribeiro-Neto, Juliana S.Teixeria. – ACM SIGMOD Record 31(2), 2002. – P. 84–93.

2 Троелсен, Э. Язык программирования C# 2010 и платформа .NET 4.0 / Э. Троелсен. – СПб. : Вильямс, 2010. – 1392 с.

3 Документация к Selenium [Электронный ресурс]. – Software-Testing.ru Россия. – Москва, 2011–2017. – Режим доступа: http://www.http://selenium2.ru/. – Дата доступа: 22.01.2017.