УДК 004.94

О МЕТОДАХ СБОРА СТАТИСТИЧЕСКОЙ ИНФОРМАЦИИ ДЛЯ АНАЛИЗА ВТОРИЧНОГО РЫНКА АВТОМОБИЛЕЙ

А. В. ПЕНЬКОВСКИЙ, В. А. ШИРОЧЕНКО
Белорусско-Российский университет
Могилев, Беларусь

Мы живем в эпоху BIG DATA, т. е. в эпоху больших массивов данных, которые уже собраны и хранятся либо в специально организованных хранилищах, доступ к которым предоставляется строго их владельцу, либо в свободном доступе – в сети Интернет.

Примером такого источника данных могут выступать сайты, на которых сами пользователи размещают объявления о продаже своих товаров. Данные сайты позволяют как продавцам, так и потенциальным покупателям предоставлять или получать информацию об одном каком-то конкретном товаре. Так, всем известный крупный интернет-ресурс для размещения объявлений о продаже Авито-Авто (www.avito.ru) [1] позволяет получить информацию о цене конкретного автомобиля, проданного на вторичном рынке, и все его параметры.

Представляет интерес изучение снижения стоимости автомобиля во времени в зависимости от различных факторов. Знание рыночных механизмов снижения цены в зависимости от года выпуска автомобиля, его пробега (заявленного, по крайней мере официально в объявлении продавцом) может быть использовано в работе эксперта – оценщика. А потребность в оценке остаточной стоимости транспортного средства возникает довольно часто.

Для анализа рынка необходимо собрать данные по предложениям на вторичном рынке автомобилей. Авито-Авто представляет собой интернет-ресурс, предназначенный для размещения предложений о продаже подержанных автомобилей. Данный интернет-ресурс имеет в распоряжении более 650 тысяч объявлений, по которым можно судить о вторичном рынке автомобилей.

В современных информационных технологиях создан и широко используется такой инструмент, как парсинг сайтов. Парсинг представляет собой процесс последовательного синтаксического анализа той информации, которая размещена на веб-страницах. Поскольку любой текст на экране монитора – это иерархически упорядоченный набор данных, который отображается с помощью языков программирования, процесс получения из этого набора информации может быть автоматизирован [2].

Для получения информации со всех объявлений о цене автомобиля на вторичном рынке, на первом этапе авторами исследования был использован инструмент Selenium WebDriver.

Selenium WebDriver, или просто WebDriver – это не имеющая пользовательского интерфейса программная библиотека, которая позволяет программным путем взаимодействовать с браузером, управлять его поведением, получать от браузера какие-то данные и заставлять браузер выполнять различные команды. Он используется для имитации деятельности браузера, что позволяет осуществлять задание критериев отбора объявлений, а также получать код HTML-страниц объявлений для дальнейшей их обработки. Для извлечения и обработки данных из HTML-документов использовался инструментарий библиотеки языка Java jsoup.

После обработки HTML-страницы, содержащей всю информацию, представленную в объявлении, инструментарием jsoup были получены необходимые для анализа данные объявления.

Далее при помощи библиотеки Hibernate для хранения и дальнейшей обработки информации была создана реляционная база данных (далее БД) MySQL. Hibernate – это библиотека, позволяющая связывать Java-классы с реляционной БД MySQL и работать с таблицами базы данных как с Java объектами [3]. Формат представления данных, после занесения в БД представлен на рис. 1.

id	age	marka	mileage	model	owners	powerOfEngine	price	privod	spaceMotor	state	transmission	typeOfCarCase	typeOfFuel
1	1	Ford	24000	EcoSport	1	122	799000	передний	1.6 л	не битый	механика	внедорожник	бензин
2	5	Renault	99000	Duster	3	135	550000	полный	2.0 л	не битый	механика	внедорожник	бензин
3	18	Volkswagen	348000	LT	4	102	340000	задний	2.5 л	не битый	механика	микроавтобус	дизель
4	5	Audi	79000	A6	2	180	1099000	передний	2.0 л	не битый	вариатор	седан	бензин
5	10	Opel	265000	Vectra	1	140	330000	передний	1.8 л	не битый	механика	универсал	бензин
6	14	Chevrolet	169000	Niva	2	80	190000	полный	1.7 л	не битый	механика	внедорожник	бензин
7	17	ВАЗ (LADA)	200000	2110	2	81	62000	передний	1.5 л	не битый	механика	седан	бензин
8	4	Renault	72500	Duster	2	102	555000	полный	1.6 л	не битый	механика	универсал	бензин
9	21	Volkswagen	220000	Polo	3	101	30000	передний	1.6 л	битый	механика	седан	бензин

Рис. 1. Формат представления данных с Авито-Авто после занесения в БД

СПИСОК ЛИТЕРАТУРЫ

1. Продажа подержанных автомобилей, купить авто с пробегом в России на Avito [Электронный ресурс]. – Режим доступа: www.avito.ru/rossiya/avtomobili/s_probegom. – Дата доступа: 28.12.2018.

2. Selenium – Web Browser Automation [Электронный ресурс]. – Режим доступа: https://docs.seleniumhq.org . – Дата доступа: 11.12.2018.

3. **Эккель, Б.** Философия Java / Б. Эккель. – Москва: Питер, 2016. – 809 с.