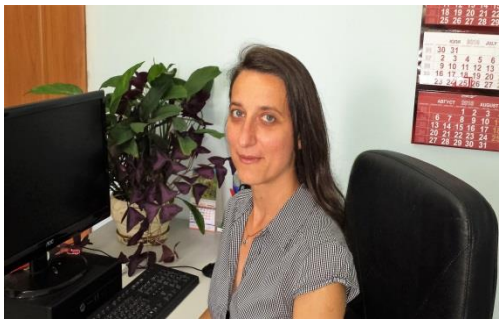
WEB SCRAPING - ИЗТОЧНИК НА ИНФОРМАЦИЯ ЗА ЦЕНИ, УЧАСТВАЩИ В ИЗЧИСЛЯВАНЕТО НА ИНДЕКСИТЕ НА ПОТРЕБИТЕЛСКИТЕ ЦЕНИ

*Десислава Захариева**



Въведение

През последните години интернет се превърна в съществена част от живота на хората и начин за комуникация между отделните институции, правителства, компании и населението. Поради това интернет пространството става все по-важно за националните статистически институти, тъй като икономическите и социалните въздействия на тази среда трябва да бъдат обхванати от статистиката. Все повече търговци на дребно се възползват от възможността да предлагат стоките си онлайн. Резервирането на самолетни билети и хотели става основно или единствено по интернет. Информацията от посочените примери може да се използва от националните статистически служби за изготвянето на официални статистически данни. Отчитайки все по-бързото разрастване на технологиите, националните статистически институти търсят нови начини за събиране на информация за статистическите изследвания. В стремежа си за иновации и използване на съществуващи източници на данни част от националните статистически служби, подкрепяни от Евростат чрез финансиране на проекти, се насочват към използването на техниката *web scraping* за получаване на информация от интернет.

Web scraping[1] е техника, с която се извлича информация от уебсайтове. Софтуерът за *web scraping* може да осъществява директен достъп до интернет (world wide web) използвайки HTTP/HTTPS протоколи или чрез уеббраузър. Терминът *web scraping* се отнася до автоматизирани процеси, които се извършват от уебробот и изключват

* Главен експерт в отдел „Потребителски цени, цени на жилища и ППС“, НСИ; e-mail: dzaharieva@nsi.bg.
[1] Статия в Уикипедия: Web scraping. (2018, Юли),Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/wiki/Web_scraping.

ръчното участие на уебпотребител при сваляне на наличната информация от сайта. *Web scraping* е форма на копиране, при която се събират конкретни данни от мрежата и се съхраняват в локална база данни или във вид на електронни таблици, с цел свалената информация по-късно да бъде използвана или анализирана.

Техниката *web scraping* се използва за извличане на информация за имейлите на потребителите на даден уебсайт, както и като компонент на приложенията, използвани за уеб индексиране; за извличане на информация от мрежата и за извличане на данни; за наблюдаване на ценовите промени онлайн и за сравняване на цените; за преглед на продуктите за наблюдаване; за събиране на обяви за недвижими имоти; за мониторинг; за откриване на промени в уебсайтове; за проучване; за проследяване на наличието на даден продукт онлайн и получаване на информация за неговия рейтинг.

Растящото значение на онлайн търговията изисква още по-голям брой цени, събирани от интернет, което натоварва допълнително експертите, занимаващи се с наблюдаване на цените. Сваляне на цените от уебсайтовете преди се извършваше ръчно, използвайки copy-paste процедура и допълнително обработване на данните, за да се достигне до подходящ вид за обработка и анализ. От друга страна, автоматичното сваляне на цени би допринесло за повишаване на качеството на данните, използвани за статистически цели.

I. Първи крачки в използването на техниката *web scraping* за автоматично сваляне на цени от интернет

Повлияна от положителния опит на останалите страни, България направи първите си крачки в използването на техниката *web scraping* за набавяне на информация за цените от интернет през 2015 г., когато започна първият пилотен проект в областта на *web scraping* техниката за сваляне на цени от интернет. Основната цел на проекта бе да се направи експеримент с автоматизирано събиране на цени за избрани продукти, които се регистрират ръчно от интернет. Специфичните цели на проекта бяха ориентирани към създаването на програма за експериментално сваляне на цени и оценка на възможните методологични и практически проблеми, свързани със събирането на детайлни цени от интернет.

Страните, чийто опит послужи за основа на първите стъпки и вземането на решение относно програмата за *web scraping,* са Австрия, Обединено кралство, Германия, Италия и Нидерландия.

Някои от страните използват готов софтуер *iMacros* (Италия), комбиниран с код, написан на *Java,* за да въведат, изберат, изтрият и сортират данните за цените на продуктите от интернет. Други страни са разработили собствена програма за *web scraping,* написана на *Python* (Обединено кралство) или *R* (Нидерландия). Различните начини за автоматично сваляне на цени от интернет си имат своите предимства и недостатъци.

Програмата import.io, която се използва от австрийската статистика, не изисква програмни умения, тъй като работи на принципа *избиране и кликване върху обекта,* който искаме да се свали (point and click web crawlers). Това позволява използването му от хора, без опит в програмирането, каквито са статистиците, които най-добре могат да преценят кои характеристики за даден продукт трябва да се изберат, за да се свалят от програмата.

Този начин на автоматично събиране на информация има и своите отрицателни страни. Ако процесът е оставен само в ръцете на статистиците, няма да може да се използва пълният потенциал на наличната технология. Това е така, защото се предпочита да се свалят само ограничен брой регистрации, както се прави при ръчното регистриране на цените, за да могат да се интегрират в разработените вече процеси за изчисляване на индекса на потребителските цени.

Създаването на собствена програма за *web scraping* използва пълния потенциал на техниката и позволява свалянето на цялата налична информация за даден сайт, структурирана според зададени критерии, но това води до определени изисквания към работещите с програмата. Те следва да имат познания в областта на програмирането, защото промяната в съдържанието на сайтовете води до промяна и в кода на програмата.

След стартирането на проекта експертите, работещи в областта на цените, направиха проучване за опита на другите страни относно процеса и въвеждането на автоматично сваляне на цените от интернет. По време на проучването се стигна до следните изводи:

- Програмата за автоматично сваляне на цени от интернет, трябва да структурира данните от страницата в подходящ формат и да сваля цялата налична информация за даден продукт.
- Програмата трябва да отчита индивидуалната структура на всеки сайт, който съдържа много специфични характеристики.
- Процесът на ръчна обработка на формата на данните, получени чрез copy-paste, може да бъде ограничен или напълно заменен чрез процедурата web scraping.

- Периодът на сваляне на данни с програма е много по-кратък и процедурата по сваляне може да бъде повтаряна многократно.

От опита на другите страни в използването на автоматично събиране на цени от интернет бяха направени обобщени изводи за предимствата и недостатъците на автоматичната процедура.

Предимства:

- Ефективността на автоматичното събиране на цени е съществено по-голяма от тази при ръчното.
- Свалянето на информацията е своевременно.
- Честотата на сваляне може да бъде повече от един път месечно.
- Получените данни са много по-обширни.
- Подобряване на качеството на изчислявания индекс чрез увеличаване на качеството на данните.
- Информацията за характеристиките на продуктите е по-подробна.

Недостатъци:

- Промяна във формата/структурата на сайта води до промяна на програмата на робота.
- Необходими са ресурси, които трябва да се инвестират в компютри и персонал със специални софтуерни умения.
- Няма информация за продадените количества на избраните продукти.

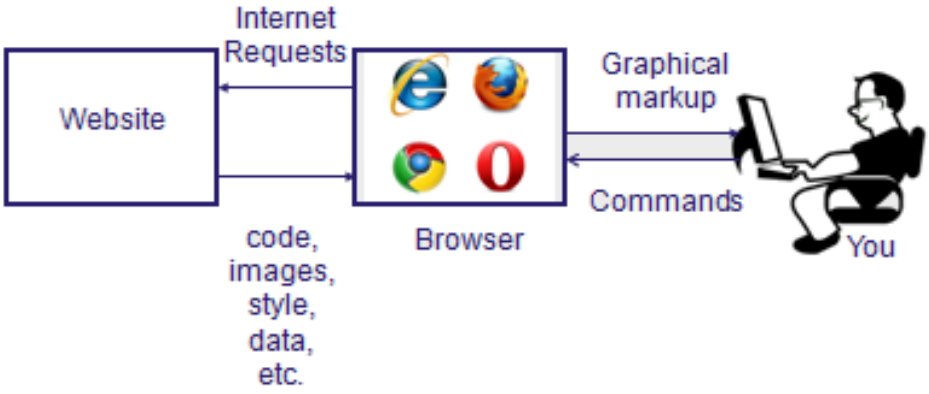
Следващата стъпка бе създаването на програма за автоматично събиране на цени от интернет, която да замени ръчното събиране на цените, да подобри качеството на данните и да позволи увеличаване на честотата на сваляне на цените за избраните продукти. Създаването на програмата породи и въпроси за законността на използването ѝ. Правните аспекти на свалянето на данни от интернет не са конкретно споменати в българското законодателство, но правата на авторите на базите данни са защитени съгласно българския Закон за авторското право и сродните му права. Експерименталното автоматизирано събиране на данни за избраните продукти и интернет сайтове, извършени в рамките на проекта, не нарушава закона поради следните причини:

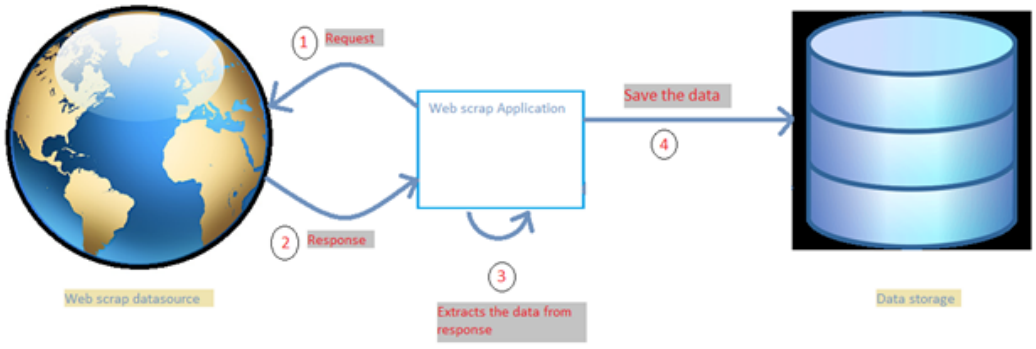
- Извлечените данни не се разглеждат като съществени части от база данни.
- Обемът на изследванията по автоматични запитвания не засяга оперативността на уебсайтовете, тъй като не е достатъчно голям.

- Извлечените данни се използват само за статистически цели.

Представената диаграма показва какъв е принципът на двата начина на регистриране на цени - ръчно и използвайки техниката *web scraping*.

Фиг. 1. Сравняване на процеса на регистриране на цени с ръчно сваляне на данни и с *web scraping*





Програмата за *web scraping*, която се използва в отдел „Потребителски цени, цени на жилища и ППС“ на Националния статистически институт, е написана на програмен език Java. Основната ѝ цел е да сваля страниците за избрани продукти, да извлича наличната информация за тези продукти и да генерира изходящи файлове в табличен

вид (csv формат), имащи определена структура на изходните таблици. Програмата сваля два основни типа страници - страница с обяви (listing page - уебстраница със списък от оферти или линкове към страници с оферти) и офертна страница (уеб-страница, в която са представени всички необходими данни за офертата). Софтуерът съхранява всички свалени страници, така че, ако има липсващи данни за параметрите в изходната таблица, съответният файл може да бъде лесно открит и да бъде анализиран за промени, което води до лесно отстраняване на възникнали проблеми.

Изводите от този експеримент са:

- Технологията за сваляне на информация от уебсайтове предлага възможност да се подобри качеството на статистическите данни и да се намали общото натоварване за събиране на данни. Тя позволява по-добро покритие на референтните продукти, повишава качеството на ценовите индекси и е много по-ефективна от гледна точка на спестяване на време.
- Ясно е, че уебсайтовете ще се оценяват за всеки отделен случай, тъй като е много трудно да се правят общи коментари за методологичните въпроси.
- Много от важните методологични въпроси могат да бъдат проучвани чрез „учене в движение“.

II. Следващи стъпки в опита с използване на *web scraping* техниката за изчисляване на индекси на цени

Следващата стъпка, използвайки опита, който придобихме и получените резултати, бе да се автоматизира процесът на регистриране на цените от свалените файлове чрез *web scraping* и въвеждането им в среда за изчисляване на индекси на цените за конкретни продукти. Обсъдени бяха няколко варианта за автоматизиране на процеса за изчисляване на индекси на цени, като едновременно с това беше разгледана и използваната практика за изчисляване на индексите в отдела.

Автоматичното изчисляване на цената, която участва в изчисляването на индекса за даден продукт, става по следната формула:

- Средната цена се изчислява като средна геометрична от всички цени (реални и импутирани) по формулата:

$$P_t = \sqrt[n]{\prod_1^n P_t^i} \ ,$$

където:

P_t е цената за текущия месец;

P_t^i е цената на *i*-тата регистрация;

n е броят регистрирани и импутирани цени.

- Изчисляване на импутирана цена

Индексът $I_{t/t-1}$, с който се извършва импутацията, се изчислява по формулата:

$$I_{t/t-1} = \sqrt[m]{\prod_1^m \frac{P_t^i}{P_{t-1}^i}} \ , \ i \in S_{t-1} \cap S_t \, ,$$

където:

P_t^i е наличните цени за периода t;

P_{t-1}^i е наличните цени (реални и импутирани) за периода t-1;

S_t е множеството от налични цени за периода t;

S_{t-1} е множеството от налични цени за периода t-1;

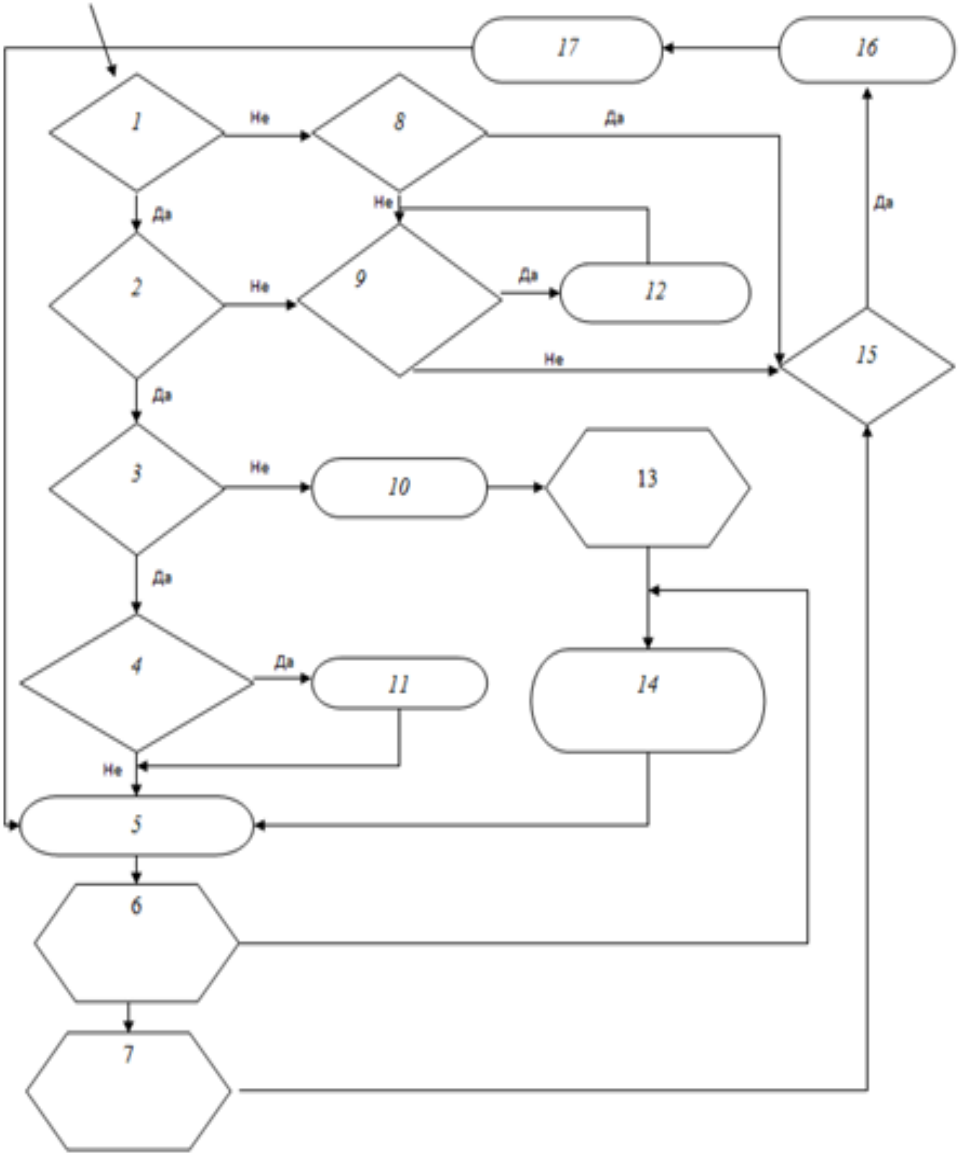
m е броят на цените, принадлежащи на подмножеството $S_{t-1} \cap S_t$.

За електронните продукти (мобилни телефони и таблети) бе решено да се направи и експериментално автоматично извеждане на предложения за замяна на липсваща регистрация на продукт, което доведе до промяна в програмата и преподреждане на изброените характеристики за различните сайтове в изходните файлове, за да могат да бъдат по-лесно обработвани и анализирани данните.

Използваната процедура за автоматизиране на процеса оцветява всички параметри, които съответстват на тези от наблюдаваната оферта. Това прави сравняването на отделните оферти много по-лесно за експерта и му позволява да вземе по-бързо и правилно решение за замяна на липсващата регистрация.

Процедурата за замяна става съгласно регламентите за ХИПЦ (хармонизиран индекс на потребителските цени), като последователността от решения и предприетите действия за замяна на дадена оферта могат да бъдат изразени чрез следната блок схема:

Процедура за замяна на обект и продукт



Легенда:

1. Отворен ли е магазинът/обектът?

2. Съгласни ли са управителите/собствениците на магазина/обекта да сътрудничат?

3. Наличен ли е наблюдаваният продукт (РО)?

4. Голяма ли е разликата с цената от предходния месец?

5. Запишете цената и попълнете специалния формуляр.

6. Ако има вероятност продуктът да липсва следващия месец, докладвайте и при необходимост пристъпете към процедура на замяна.

7. Ако има вероятност обектът да бъде затворен през следващия месец, докладвайте и при необходимост пристъпете към процедура на замяна.

8. За постоянно ли е затворен магазинът/обектът?

9. Отказът само на отсъствие на лицето за контакт ли се дължи?

10. Докладвайте за липсата и причината за нея.

11. Обяснете.

12. Опитайте отново. В случай на провал докладвайте фактите.

13. Ако наблюдаваният продукт (РО) липсва за втори пореден месец.

14. Изберете нов продукт, който по своята същност е еднакъв с липсващия, запишете характеристиките му и попитайте за цената му от предходния месец.

15. Избор от регистратора на цени?

16. Изберете нов обект съгласно инструкциите.

17. Изберете продукт, който по своята същност е равностоен/еднакъв с наблюдавания.

Осъзнавайки факта, че използвайки получените данни, за да се прилага методологията за единично регистриране на цени, не се оползотворява напълно техният потенциал, експертите стигнаха до решението да се експериментира с нови начини за изчисляване на индексите на цените и да се предприемат стъпки за увеличаване на периодичността на сваляне на цените от интернет за избрани продукти в даден месец. Може би по-важното е, че *web scraping* ни дава възможност да изследваме големи масиви от данни и да експериментираме с методологии, които са подходящи за обема на данните. Този опит ще бъде от полза, ако в статистиката на потребителските цени се въведат други източници на информация за изчисляване на индекса (например сканирани данни[2]).

Опитът на страни, които прилагат частично или напълно потенциала на данните, получени с *web scraping,* показва, че от методологична гледна точка за изчислението на индексите могат да се използват препоръчаните от Евростат методи, които се прилагат за изчисляването на индекси от сканирани данни.

За разлика от сканираните данни данните, получени чрез *web scraping,* не съдържат информация за направените разходи. Това означава, че индексите не могат да бъдат претеглени на по-ниско ниво от нивото на групите елементарни агрегати. Освен това при традиционното регистриране на цени регистраторите избират продукти, които обикновено се купуват от потребителя (този избор обаче е повече или по-малко субективен, тъй като той не се основава на емпирични данни за най-продавания продукт от съответната категория (елементарен агрегат на ECOICOP - Европейска класификация на индивидуалното потребление по цели), а на личната преценка на

[2] Сканираните данни се генерират от касовите устройства в магазините и представляват информация за оборота на магазина - количествата продадени стоки по GTIN (Global Trade Item Number), познат в миналото като EAN номер или, просто казано, като баркод за определен период.

регистратора), докато при данните, получени чрез *web scraping*, експертите ще изберат всички продукти независимо от направените разходи.

Заключение

Разработването и тестването на техниката *web scraping* за целите на потребителските цени показа огромната възможност за намиране на цени на продукти и свързаните с тях характеристики в мрежата, но поставя и много важни въпроси от гледна точка на методологията и използването на тези данни за статистически цели.

От гледна точка на ефективността на процеса за регистриране на цени от интернет, експериментите, направени с няколко продукта, доказаха ефективността на техниката *web scraping*.

В заключение може да се каже, че използването на web scraping техниката в областта на потребителските цени като източник на информация има своето логическо място сред източниците, участващи в изчисляването на индексите на потребителските цени.

ЦИТИРАНА ЛИТЕРАТУРА:

Boettcher, I. (2015). Automatic data collection on the Internet (web scraping), Statistics Austria, Paper presented on the 2015 Ottawa Group on Price Indices.

Giannini, R., R. Lo Conte, S. Mosca, F. Polidoro, F. Rossetti (2014). Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation, ISTAT, European Conference on quality in official statistics, Vienna, 2-5 June 2014.

Hoekstra, R., O. Bosch, F. Harteveld (2010). Automated Data Collection from Web Sources for Official Statistics: First Experiences, Statistics Netherlands.