ПРИЛОЖНИ АСПЕКТИ НА „ГОЛЕМИТЕ ДАННИ“ В ОФИЦИАЛНАТА СТАТИСТИКА

*Галя Статева**

Въведение

Процесите на глобализация и технологизация във всички сфери на обществения живот оказват огромен натиск по отношение на управлението в национален и международен аспект. За успешното управление е необходима „добра“ информация, което означава, тя да притежава следните характеристики: синтезирана, кратка, точна, акцентирана, подсказваща най-доброто решение на проблемите и очертаваща (прогнозираща) хоризонтите за развитие на процесите и явленията. В този смисъл Big Data е друга, различна от тази, която познаваме, екосистема. Анализът на основата на „големите данни“ може да бъде отвъд източниците, измерването и напрежението, необходими при създаването на информация, а също така отвъд политиката. Успоредно с това през 21-ви век тази екосистема е исторически феномен на човешкото развитие.

В последните няколко години редица национални статистически офиси осъществяват европейски и международни проекти за „големите данни“. Опитът от успешното използване на „големите данни“ може да бъде изучаван и споделян с други държави с цел извличане на ценни познания и прилагане на добри практики по отношение на Big Data. Освен това националните статистически организации са окуражавани от Евростат да включат официално въпросите и за

---

* Д-р, държавен експерт в отдел „Обща методология и анализ на статистическите изследвания“, дирекция „Методологично-учебен център“, Национален статистически институт; e-mail: GStateva@nsi.bg.

„големите данни“ в техните годишни програми и стратегически документи чрез осъществяване на експериментално-изследователски и пилотни проекти в избрани области и чрез разпределяне на подходящите ресурси за тези цели. В процеса на приложение на Big Data в официалната статистическа практика Евростат играе ключова роля, като се има предвид факта, че той е източник на финансови, идейни и технически ресурси, включително и организатор на редица специализирани обучения, семинари и научни конференции по тази тематика.

Националните статистически институти създават разнообразни масиви от статистически данни за използване на информационните и комуникационните технологии, които се използват за наблюдение на напредъка на страните към информационното общество. Традиционно тези данни се събират чрез два различни въпросника - един за домакинствата/физическите лица и един за предприятията. Бързото развитие на съвременните ИКТ поставя необходимостта от разработване на показатели, които да са повече релевантни и навременни от тези, изчислявани на базата на традиционните изследвания.

Дигиталните отпечатъци/следи, резултат от нашето ежедневие, могат ефективно да бъдат използвани за измерване на голямо разнообразие от явления. В проведено проучване през 2013 г. Европейската комисия изучава възможността за използване на интернет като допълнителен източник на данни или дори като заместител на традиционните статистически източници[1]. Като резултат от изследването се стига до заключението, че относно текущите променливи за физически лица/домакинства нито достъпът до ИКТ, нито използването на компютри са възможни за измерване чрез интернет. От друга страна, прогнозируемото използване на интернет мрежата от предприятията е идеално за измерване чрез интернет. Налага се мнението, че събраните данни по електронен път биха предоставили достатъчно материал за изчисляване на допълнителни индикатори за ИКТ, както и голям брой потенциални показатели, които предлагат допълнителни възможности за изучаване на съществуващи явления и необхванати в досегашното традиционно статистическо изследване за ИКТ в предприятията.

През 2013 г. Италианският статистически институт (ISTAT) започва да тества техники за извличане на информация от интернет (web-scraping) и извличане на съдържание от текст (text mining) за около 5 600 уебсайта от общо 8 600 адреса на уебсайтове, посочени от предприятията при

[1] Internet as data source. Feasibility Study on Statistical Methods on Internet as a Source of Data Gathering. (2012). European Commission, SMART 2010/30. file:///D:/Fullfinalreportofthestudy.pdf.

попълване на въпросника на традиционното изследване, като събраното съдържание се индексира чрез технологични инструменти и алгоритми и се оценяват резултатите от статистическото изследване. По този начин ISTAT прави практически опит за промяна на традиционното изследване „Използване на ИКТ в предприятията“ (Information and Communication Technologies in Enterprises)[2] чрез осигуряване на допълнителни източници на данни. Заключенията и съображенията от опита на ISTAT относно използването на интернет като източник на данни за официалната статистика са проучени от българските специалисти и са основа, на която те проектират настоящото изследване.

В началото на 2016 г. Националният статистически институт на Р България (НСИ) се включи като страна партньор в ESSnet проект „Рамково споразумение за сътрудничество Big Data План за действие“, който е разделен на две отделни грантови споразумения (SGA-I и SGA-II), изпълняващи се последователно за периода 2016 - 2018 година. В рамките на това споразумение се постигат целите, заложени в т.нар. BDAR (Big Data Action Plan and Roadmap v. 1.0), който е част от портфолиото на ECC Vision 2020. Основната цел на проекта е да подготви Европейската статистическа система за интегриране на източници на „големи данни“ в процеса на производство на официална статистика. Специфичните цели на работните пакети са свързани с анализиране на получените резултати от източниците на Big Data; разработване на методология за използването на Big Data в статистическата практика и измерване на качеството на събраните данни; идентифициране, дефиниране и внедряване на ИТ инфраструктура за обработка и съхранение на Big Data; правни въпроси, свързани с достъпа и използването на източниците на Big Data в рамките на ECC; обмен на информация между официалната статистическата система и научната общност.

Статистиката за използването на ИКТ е естествен и логичен „кандидат“ за пилотен проект и реинженеринг на базата на интернет и подобни източници. Поради тази причина **през 2016 г. екип[3] от НСИ** провежда емпирично изследване на тема „Извличане на информация от интернет за характеристики на предприятията (web-scraping)“ в рамките на европейския проект.

Основната цел на проведеното емпирично изследване е аналогична на проведеното от ISTAT изследване и е насочена към проучване на възможностите за прилагането на техниките „web-scraping“ и „text mining“ и други подобни, както и да се оцени ефектът от използването им в процеса на събиране на данни и подобряване на качеството на информацията за предприятията от бизнес

[2] Information and communication technologies in enterprises. http://www.istat.it/en/archive/77760.
[3] Авторът на статията е ръководител на екипа на НСИ и участва активно като основен статистически експерт при организацията, провеждането и анализа на резултатите от изследването.

регистъра на НСИ чрез достъп до техните уебсайтове. Специфичните цели на изследването са: да се демонстрира дали статистическите бизнес регистри могат да бъдат подобрени чрез използване на web-scraping и прилагане на моделно-ориентирани подходи, за да се предскажат стойностите на някои ключови променливи за всяко предприятие; да се верифицира възможността за производство на статистически резултати от масив с големи данни с по-надеждна прогнозна стойност и съчетаването им с данни от традиционно статистическо изследване или административни данни. За сравнение и валидиране се използват данните, получени чрез статистическото изследване „Използване на ИКТ от предприятията“, което се провежда регулярно и съгласно европейските стандарти и националното законодателство от НСИ. Наблюдението е годишно, извадково, като в обхвата му на случаен принцип се включват около 4 900 предприятия. Генералната съвкупност обхваща всички предприятия от нефинансовия сектор с 10 и повече заети лица, включени в статистическия бизнес регистър на НСИ. Именно тази съвкупност е базова за провеждане на изследването за извличане на „големи данни“ от интернет чрез прилагане на техниките „web-scraping“. В списъка са включени 26 836 предприятия, които са проучени за наличие на сайтове. Първоначално в записите на бизнес регистъра са открити 2 006 броя фирмени URL адреси и 20 649 броя имейл адреси.

Работният процес на изследването е организиран в *четири основни фази*, а именно: фаза 1: Спецификация на т.нар. „сценарии“ („use-cases“), за да се дефинира обхватът на изследването; фаза 2: Разработване на един или повече пилотни проекта за всеки „use-case“, които да бъдат внедрени експериментално в НСИ; фаза 3: Практическа реализация на пилотните проекти от НСИ и фаза 4: Идентифициране на основните методологически и технологични въпроси като резултат от реализацията на пилотните проекти.

За постигане на специфичните цели на емпиричното изследване са осъществени няколко основни задачи, както следва:

Задача 1: Достъп до данни. В рамките на тази задача са извършени дейности по идентифициране на набор от методи за търсене на URL адреси за предприятия, за които те не са налични и проучване на правните аспекти относно достъпа до данни на уебсайтовете на предприятията. Проучването на правните и етичните аспекти за достъпа и съхранението на „големи данни“ беше особено важна предпоставка и условие за стартиране на действителната работа по проекта. Законодателната рамка за използване на данни от други източници за производство на

официална статистика в България се състои от Закона за статистиката, Закона за защита на личните данни, Закона за авторското право и сродните му права, Закон за електронната търговия, Общ регламент за защита на данните (приложим в България от 25 май 2018 г.) и не представлява пречка за извършване на дейности по „извличане“ на информация от Интранет мрежата. По отношение на надлежно осведомяване на предприятията, които са обект на изучаване в контекста на извличане на данни от корпоративните им уебсайтове, НСИ предприе специална кампания на сайта си, за да информира всички заинтересовани предприятия и други потребители, че ще бъде извършен масивен „web-scraping“ с експериментална цел за нуждите на официалната статистика.

Задача 2: Боравене с данните. В обхвата на тази задача се включва подробното дефиниране на „сценариите“, провеждане на консултации с потребители и други заинтересовани лица за анализиране на нуждите от статистически данни и координация с дейностите по проект ESS.VIP „Европейска система за оперативно съвместими бизнес регистри“. Проучена е научната литературата относно техниките и наличните свободни софтуерни продукти за масивен „web-scraping“ (JSoup, HTTrack и т.н.) и изучаване на проблемите, свързани с достъпността на уебсайтовете (блокиращ механизъм). Предвижда се внедряване на една или повече техники за „web-scraping“ и провеждане на същинското извличане на съдържание на уебсайтовете на предприятията и последващото им съхранение на събраното съдържание в база данни. За дефиниране на обхвата на изследването са избрани следните **use-cases:**

- **Use-case 1.** Генериране на списък с фирмени URL адреси на предприятията за бизнес регистъра **(URLs retrieval).**
- **Use-case 2.** Електронна търговия в предприятията (E-commerce) - прогнозиране дали дадено предприятие предоставя възможности за електронна търговия на фирмения си уебсайт, или не.
- **Use-case 3.** Присъствие на предприятията в социалните медии (Social media presence) - търсене и събиране на информация от фирмения уебсайт дали дадено предприятие съществува в различни социални медии.
- **Use-case 4.** Апробиране на софтуер в НСИ (разработен от италианския статистически офис ISTAT) за генериране на списък с фирмени URL адреси на предприятията (URLs retrieval) и сравнение на получените резултати от българския и италианския софтуер.

Задача 3: Тестване на методи и техники. За да се приложат на практика техниките „web-scraping“ е необходимо разработването на софтуерни инструменти, които се прилагат в съответствие с дефинираните „сценарии“. В рамките на тази задача екипът е разработил авторски софтуер, чрез който практически е извършено изследването. На следващо място е осигуряването на информация с цел обогатяване на статистическия бизнес регистър и изграждане на доказателства за избраните „сценарии“, за да се прогнозират характеристики на предприятията чрез прилагане на „text mining“ и „data mining“ към събраните URL на предприятията.

Задача 4: Финализиране на методи и техники. Тази задача включва: избор на извадка от уебсайтове и ръчно определяне на свързани характеристики на предприятия и/или използване на резултатите от изследването „Използване на ИКТ от предприятията“, за да се валидират някои характеристики на предприятията; прилагане на техники за извличане на текст и данни, за да се прогнозират характеристики на предприятията; оценяване на „прогнозираните“ характеристики на предприятията чрез показатели за качество (например точност, чувствителност, специфичност) и сравняване и евентуално интегриране на получената информация с информацията от националните бизнес регистри.

За нуждите на изследването беше необходимо да се направи *концептуално сравнение* между етапите на класическия производствен процес за провеждане на едно традиционно изследване и основните фази на бизнес процеса за получаване на информация от източници на Big Data. За тази цел беше използвана националната версия на **Общия модел на статистическия производствен процес (ОМСПП)** в съответствие с GSBPM (Generic Statistical Business Process Model ver. 5.0)[4], който е адаптиран към потребностите на българската статистическа система и е наличен на сайта на НСИ (http://www.nsi.bg/sites/default/files/files/metadata/NSI_GSBPM_2016.pdf.). За разлика от класическото статистическо изследване процесът на получаване на информация от източници на Big Data е съществено различен, тъй като източниците на данни, достъпът до тях и средствата за тяхното събиране са напълно различни. В този смисъл може да се твърди, че GSBPM моделът е неприложим в настоящия си вид за описание на процеса на работа с Big Data. Въпреки това някои от гореописаните етапи на работа за „извличане“ на информация от интернет **могат да се съотнесат** (с определена условност) към фазите и подфазите на ОМСПП. *Първоначалното дефиниране на „сценарии“* и тяхното съдържание е съотносимо с фаза 2 от ОМСПП „Проектиране“; *задачите, свързани с достъпа до данни* от интернет източниците, съответстват на фаза 4 от ОМСПП

[4] Generic Statistical Business Process Model, v5.0, (2013). https://statswiki.unece.org/display/GSBPM/GSBPM+v5.0.

„Събиране на данни“, която включва събирането на първичните статистически данни и зареждането им в подходяща ИТ среда за по-нататъшна обработка в контекста на традиционното статистическо изследване; *боравенето с данните от източници на Big Data* може да се свърже частично с фаза 4 и изцяло с фаза 5 от ОМСПП „Обработка на данни“, където с различни ИТ средства и техники се обработват и редактират вече събраните първични данни, като се съхраняват в подходящи бази данни; *тестването на методи и техники по отношение на вече събраните и обработени Big Data* отговаря частично на някои подфази на фаза 5 и изцяло на фаза 6 „Анализ“ от ОМСПП.

След първия опит в събирането на данни от интернет е необходимо уточнението, че съществува разграничение между автоматизираното събиране на данни (от интернет сайтове с интернет роботи, „паяци“ и други подобни инструменти) без човешка намеса и асистирано събиране на данни (подпомагано от оператори). За втората категория е важно да се подпомага операторът, който събира данните, с цел проверка за настъпили промени в данните, достъпни на интернет сайтовете. Трябва да се отбележи обаче, че и двете категории за събиране на данни от интернет са значими и с положителни възможности за официалната статистика, включително за изучаване на социално-икономическите явления по съвсем нов начин. Автоматизираното събиране на данни може да доведе до по-детайлни данни в сравнение с данните от статистическото изследване, които могат да се използват за потвърждаване на работата на статистическите експерти, за подобряване на ефективността или за намаляване на тежестта на респондентите. От своя страна, асистираното събиране на данни би било полезно за събиране на цени на стоки и услуги от много на брой интернет сайтове по значително ефективен начин.

Идентификацията на уебсайта е друг важен въпрос, когато се използват уебсайтове за събиране на данни. Необходимо е да се оцени надеждността на сайтовете съобразно обекта на изследване, колко лесно четими са данните, кои променливи са налични и колко сравними са те сред данните от различни сайтове. Освен това трябва да се знае как нараства обемът на данните и каква е тяхната променливост - все типични характеристики за големите данни, които се събират от интернет пространството. За разлика от традиционните източници на данни (статистически въпросници, административни източници), при които характеристиките на данните са известни на организацията-доставчик или са контролирани от статистическите офиси, данните от интернет източниците са изцяло извън техния контрол.

Важен въпрос е управлението на грешките, тъй като както в традиционното изследване, така и при работа с Big Data могат да възникнат грешки на различни етапи. Някои видове грешки могат да бъдат специфични за източника; други биха могли да се прилагат за всички източници (т.е. конструктивна валидност, грешка на обхвата, грешка при измерването, грешка, дължаща се на импутация, грешка при липса на отговор от респондентите и т.н.). Извадковата грешка би могла да се приложи за специфичните случаи, където се използват извадкови методи и техники. При работа с „големи данни“ етапите за обработката им трябва да включват предварителна функция за получаване, при която данните първо се верифицират и предварително се третират, преди да последва по-задълбочен анализ, при който грешните данни, екстремалните и липсващите стойности се маркират за последваща обработка. Всички видове източници на „големи данни“ могат потенциално да са обременени от частична липса на отговор (липсващи стойности за специфични променливи). Така че в процеса за „почистване“ на „големите данни“, е необходимо да се вземе предвид, че знанието за данните и свързаните с тях метаданни са ключов фактор за разработването на ефективни методи за обработка. Предвид огромния обем на данните екстремалните стойности може да нямат влияние в сравнение с традиционната статистическа обработка. Докато методите за импутиране на данни са добре известни за традиционната статистика, все още има само няколко опита за импутиране на текстови стрингове или друг вид неструктурирани данни.

И не на последно място, огромното увеличаване на наличността на неструктурирани текстови данни изисква официалните статистически институции да увеличат финансовите си инвестиции в инструменти, способни да анализират текстови данни от интернет източници. Тези инструменти, като се започне от извличането на данни от мрежата до текстовия анализ и извличане на съдържание от текста, трябва да станат част от стандартния набор от инструменти на статистиците и анализаторите на данни.

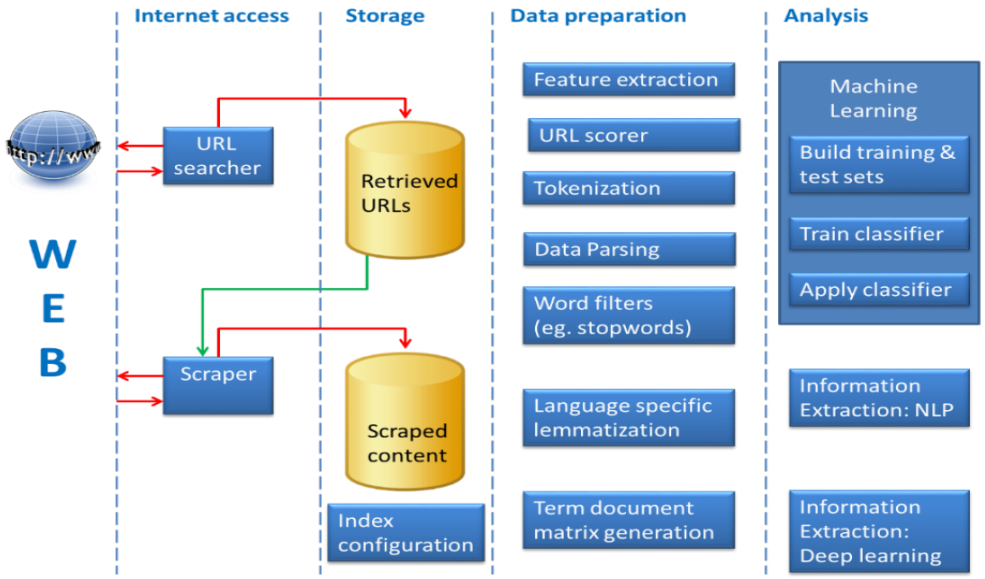
I. Технологична среда за приложението на „web-scraping“

За изпълнение на дефинираните задачи на настоящото емперично изследване и прилагане на техниките на „web-scraping“ е разработена обща референтна логическа архитектура[5], съставена от четири блока, съответстващи на четирите основни етапа на работата по извличането на данни от уебсайтовете на предприятия, а именно: „Интернет достъп“ (Internet access); „Съхранение“ (Storage);

[5] Web-scraping: Applications and Tools. (2015). European public sector information platform, Topics report № 10. https://www.europeandataportal.eu/sites/default/files/2015_web_scraping_applications_and_tools.pdf.

„Подготовка на данни“ (Data preparation) и „Анализ“ (Analysis). За всеки етап се описват логическите функционалности, които трябва да бъдат изпълнени от специфичните софтуерни продукти, разработени за нуждите на настоящото емпирично изследване (фиг. 1).

Фиг. 1. Референтна логическа архитектура на технологичния процес



В етапа **Интернет достъп** се изпълняват две логически функционалности - „URL Searcher“ и „Scraper“: функционалността на блока **URL Searcher** е да намери и състави списък от уебсайтове, свързани с дадено предприятие. Обикновено този списък се получава чрез заявка в интерент търсачка, използвайки името на предприятието като „дума за търсене“. Основното допускане е, че ако предприятието има официален уебсайт, то той трябва да се намери в рамките на резултатите, генерирани от уебтърсачката. Блокът **Scraper** е отговорен за придобиване на наличното съдържание за всеки URL адрес в списъка с URL адреси, предоставени като вход. Той може да има допълнителни функции като филтриране на URL адреси (ако е предоставен списък с такива) и обикновено е конфигурируем чрез задаване на различни параметри като дефиниране ниво на извличане на данните (само началната страница или началната страница плюс първо ниво и т.н.).

Вторият етап **Съхранение** съдържа три функционалности „Retrieved URLs“, „Scraped content“ и „Index configuration“: блокът **Retrieved URLs** е основен контейнер с URL адреси (намерени като резултат от предишната стъпка), ранжирани от обикновен файл до система за управление на бази

данни. Блокът **Scraped content** е контейнер със съдържание, получено от работата на блока „Scraper“. Обикновено е необходимо този блок да се реализира, прилагайки нетривиални решения, поради факта, че извлеченото количество информация е огромно и е съставено от неструктурирани данни. Блокът **Index configuration** представлява стратегия за индексиране на извлечените данни, съхранявани в блока „Scraped content“. В контекста на Big Data съдържанието, съхраненото огромно количество данни е параметър, който трябва да бъде взет под особено внимание, тъй като индексирането на данните дава възможност по-лесно да се открие информация в следващите фази.

Третият етап **Подготовка на данни** обхваща седем функционалности „Feature extraction“, „URL scorer“, „Tokenization“, „Data parsing“, „Word filters“, „Language specific lemmatization“ и „Term document matrix generation“. Блокът **Feature extraction** е отговорен за локализиране и намиране от извлечените от интернет данни на набор от предварително дефинирани характеристики на предприятията (например адреси, телефонни номера, имена, ДДС кодове и други). Обикновено функциите на този блок се изпълняват в специфична софтуерна програма. Блокът **URL scorer** се използва за оценяване на даден URL на базата на някои предварително дефинирани параметри, като наличието на някои характеристики, присъстващи в съдържанието на URL адреса. Като се има предвид, че в предишните стъпки е намерен определен списък с URL адреси, свързани с едно предприятие, този блок може да се използва самостоятелно или съвместно с друг блок, за да се идентифицира най-вероятният официален корпоративен URL адрес за конкретното предприятие. Блокът **Tokenization** обработва текстовото съдържание на извлечените ресурси, като го трансформира в текст, който става вход за следващия етап, например като синтактичен анализ (parsing) и „text mining“ или за етапа „Анализ“. Блокът **Data parsing** се фокусира върху анализа на „символи“, произведени от блок „Tokenization“ чрез търсене на специфични регулярни изрази, съвпадащи изречения и т.н. Блокът **Word filters** се използва за филтриране на някои думи/символи (ако е предоставен списък от думи, които трябва да бъдат филтрирани) от „извлеченото“ текстово съдържание или за обогатяването на това съдържание със списък от водещи думи. Блокът **Language specific lemmatization** извършва т.нар. „лематизация“ на символите. В конкретния случай (компютърна лингвистика) лематизацията е алгоритмичен процес на определяне на лема на дума, основана на нейното първоначално значение. Когато не е възможно да се изведе първоначалното значение, обикновено се използва основната форма на символа, получен чрез прилагане на стриминг (автоматизиран процес, който произвежда основен стринг в опит да представя семантично свързани думи), който изчислява базовата форма на символа чрез работа върху една дума, без да е познато

съдържанието. Блокът **Term document matrix generation** е отговорен за създаването на терминологично-документална матрица, която да бъде използвана от блоковете в следващия етап „Анализ“. Обикновено всяка клетка на матрицата съдържа броя на случаите на даден символ в уебсайта на предприятието.

В етапа **Анализ** функционалностите са три „Machine learning“, „Information extraction: NLP“ и „Information extraction: Deep learning“: Блокът **Machine learning** (и неговите подблокове) генерират статистически изходи чрез използване на класификатор или „обучител“ в контекста на компютърната лингвистика. Блокът **Information extraction: NLP** генерира статистически изходи чрез използване на подходи за компютърна обработка на естествен език (NLP). Блокът **Information extraction: Deep learning** генерира статистически изходи чрез използване на техники за дълбоко изучаване в контекста на подходите на изкуствен интелект.

Ключов въпрос при приложението на техниките на „web-scraping“ е дали структурата на съдържанието на уебсайта е известна предварително и могат ли да се правят различни допускания за структурата на данните преди да се извърши същинското „извличане“ на информация, или това не е възможно. Първият случай предполага използването на ***специфичен „web scraping“,*** а вторият - ***генеричен „web scraping“.***[6]

Специфичен „web-scraping“ се прилага, когато са добре познати както структурата, така и (типът) съдържанието на уебсайтовете, от които ще бъдат извличани данни. В този случай компютърните роботи възпроизвеждат човешкото поведение, като посещават автоматично уебсайта на дадена фирма с цел събиране на информация, която представлява изследователски интерес. Типични примери за такъв вид „web-scraping“ могат да се открият в областта на статистиката на цените, където повечето от позициите в онлайн магазина имат един и същ продуктов списък или продуктова страница. По този начин софтуерът за „извличане“ може да открива специфични елементи като: първоначална цена, продажна цена, етикет на стоката, описание, количество, цвят, размер и т.н. от много на брой уебстраници за много продукти.

Генеричен „web scraping“ се прилага, когато липсват априори познания за структурата и съдържанието на фирмения уебсайт и е необходимо цялото му съдържание да бъде извлечено и

[6] Boeing, G., Waddell, P. (2016). New Insights into Rental Housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listings. Journal of Planning Education and Research, 23 august. http://journals.sagepub.com/doi/pdf/10.1177/0739456X16664789; Vargiu, E., Urru, M. (2013). Exploiting web scraping in a collaborative filtering- based approach to web advertising. Artificial Intelligence Research, vol. 2, № 1. http://www.sciedu.ca/journal/index.php/air/article/view/1390.

обработено, за да се събере някаква информация, която представлява интерес за изследователя. Типичен пример за такъв вид „web-scraping“ е намирането на уебстраници на предприятията за извличане на някои общи характеристики на съвкупността. В този случай е необходимо да се разработи по-обща методология за „web-scraping“, както и за прилагане на специализиран софтуер за извличане и обработка на данните.

Разграничението и съпоставянето на двата вида „web-scraping“ е необходимо, тъй като при тяхното практическо приложение има технически и методологически разлики. Колкото по-малко се познава структурата на обекта, за който ще се прилага „web-scraping“, толкова по-общи са технологичните и методологичните подходи, които ще се използват на практика.

От технологична гледна точка със ***специфичния „web-scraping“*** е възможно предварително да се проучи обектът (уебсайтове или страници) и да се проектира специфичен „софтуер-скрапер“, който използва това знание, за да се навигира в сайта чрез идентификатори на структурата на страниците като например „*html id“*, „*xpaths“* и „*css“* селектори. С ***генеричния „web-scraping“*** обикновено се отстраняват всички html маркировки и се прилага софтуер за извличане на съдържание от текст и документи върху останалото съдържание.

От методологична гледна точка със ***специфичния „web-scraping“*** се постигат добре дефинирани променливи, които да се използват при обработката на данните. Трябва да се има предвид, че при извличането винаги има известна степен на несигурност по отношение на емпиричните данни. При ***генеричния „web-scraping“*** обикновено се прилагат машинно самообучителни методи върху резултатите от стъпките за обработката на данни.

Техниката „web-scraping“ се използва от началото на съществуването на интернет, но търпи развитие през последните години. Прокси сървърите извличат съдържанието на уебсайтове и ги съхраняват за обслужване на локални компютри преди повече от 20 години. Тази дълга история на „web-scraping“ е довела до наличието на различни инструменти и методи, които могат да бъдат използвани за събиране на информация от уебсайтове, както и техники за анализ на извлечената информация. Независимо от тяхното разнообразие те могат да се обединят в две групи: **подходи за машинно самообучение (machine learning) и детерминистични подходи.**[7] За да се обоснове изборът на съответните методи за приложението им в пилотните проекти за дефинираните „use-cases“, е необходимо да се изясни накратко тяхната същност.

[7] Alpaydın, E. (2004). Introduction to Machine Learning (Adaptive Computation and Machine Learning). MIT Press, ISBN 0-262-01211-1; Witten, I., Frank, E., Hall, M. (2011). Data Mining: Practical machine learning tools and techniques. 3rd ed. Morgan Kaufmann. ISBN-13: 978-0123748560.

В контекста на официалната статистика извличането на данни *чрез „machine learning“ подходи* се състои в прилагане на алгоритми или модели, производни от набор от данни за самообучение, които се предполага, че са представителни за даден изучаван проблем. Параметрите на модела обикновено се настройват с набор за валидиране, преди да се измери неговата ефективност на т.нар. тестови набор от данни с известни характеристики. Накрая моделът се прилага към други набори от данни, за които не е известно нищо, но за които се предполага, че моделът работи добре, за да се произвеждат статистически данни. От друга страна, извличането на данни за официалната статистика *чрез детерминистични подходи* се състои в прилагане на алгоритми, проектирани от набор от правила с известни предварително характеристики на уебсайтовете и структурата на данните. Казано по друг начин - знанието на експерта се използва за проектиране на алгоритъм за обработка и интерпретиране на входни данни от уебпространството и други източници в статистически целеви променливи. Този метод се нарича детерминистичен, защото алгоритъмът, приложен към едни и същи данни, винаги ще има един и същ (детерминистичен) резултат в сравнение с machine learning подхода, който зависи в голяма степен от обучителния набор от данни, който е използван.

Като цяло превръщането на данните от интернет в статистически данни обикновено изисква много стъпки, като във всяка стъпка може да се приложи различен подход. Един от факторите, които влияят върху избора, е сложността на взаимовръзката между входните променливи или производните характеристики, получени от входните данни и статистическите целеви променливи. Ако тази връзка е сравнително ясна, детерминистичният подход е подходящото решение и обратно, ако тази връзка е комплексна, неизвестна или трудна за моделиране в алгоритъм, което обичайно се случва при работа с уебданни, подходът за машинно самообучение е правилният избор.

Важно е да се отбележи, че при подходите за машинно самообучение наличието на набор от данни за обучение със задоволително качество е от съществено значение. В много случаи това е предизвикателство. В някои от пилотните проекти тези данни за обучение са налични или могат да бъдат получени от предишни статистически изследвания. Това би могло да е валидно в краткосрочен план, когато в официалната статистика е въведен machine learning подходът, с цел да замени (частично) традиционния статистически процес, но в дългосрочен план това почти не е възможно. Очевидно е, че детерминистичните подходи нямат това предизвикателство, но имат други проблемни въпроси, които ги правят не толкова универсално приложими. При подхода на машинното самообучение от решаващо значение е превръщането на текст, събран чрез генеричен

„web-scraping“ в модел. HTML и друга информация, съдържаща се в таговете и изображенията, е неструктурирана и съдържа много шум, който, ако не е филтриран, би направил сигналите неразбираеми. Поради тази причина трябва да се прилагат едновременно техники за извличане на съдържание от текст и данни, за да се структурират и стандартизират данните и да се открива съответната информация.

Какви методи и техники са използвани при реализацията на „use-cases“, дефинирани в настоящото емпирично изследване? В по-голямата си част събраните текстове са обработени чрез последователно изпълнение на следните стъпки: а) нормализиране: стеминг и лематизация; б) подбор на характеристики. За всяка зависима променлива (електронна търговия, присъствие в социалните медии) всички нормализирани термини са обработени, за да се открие най-подходящият термин за прогнозиране. Използвани са редица техники, като кореспондиращ анализ, регуларизационни техники (LASSO, Ridge, Elastic Net) - част от алгоритмите за машинно обучение. Целта е да се намали броят на термините до управляем брой значими независими променливи, които са използвани като вход към модели, пасващи в дадени обучителни набори от данни (получени по различни начини: чрез ad-hoc анализ на редица случаи или чрез използване на данни от регулярното статистическо изследване „Използване на ИКТ от предприятията“). По принцип обучителният набор от данни е разделен на отделни водещи елементи, като тестови набори от данни, които се използват за оценка на модела чрез сравняване на наблюдаваните и прогнозираните стойности.

Първият „use-case“ „Генериране на списък с URL адреси на предприятията“ се изпълнява на три стъпки: получаване на изходни точки от интернет търсачка; извършване на web-scraping на намерените URL адреси или друга извлечена от уебмрежата информация за тях; определяне кой от получените URL адреси като резултати от предходните две стъпки е истинският уебсайт на дадено предприятие. Първата стъпка се извършва или чрез структуриран приложно-програмен интерфейс API, или чрез извличане на уебстраница като резултат от търсещата машина. И двата случая са характерни примери за *специфичен „web-scraping“ с прилагане на техники за детерминистичен анализ*, т.е. резултатите от търсенето могат да се разглеждат като структурирано съдържание. Втората стъпка е пример за *генеричен „web-scraping“*, тъй като предварително не е известно нищо за обекта, който ще бъде подложен на изследване. При третата стъпка - фазата на анализа, се използват подходи за машинно самообучение, за да се определи коректният URL адрес от всички намерени URL адреси, както и ръчно валидиране на резултатите.

Вторият „use-case“ „Електронна търговия в предприятията“ е по същество *генеричен „web-scraping“*, при който изходната точка е списък с фирмени URL адреси, за които предварително не е известно нищо за структурата на уебсайта. Използват се подходите на *машинно самообучение* за идентифициране на електронната търговия, като едновременно с това изследването „Използване на ИКТ от предприятията“ се използва като основно за набор от данни за обучение.

Третият „use-case“ „Присъствие на предприятията в социални медии“ се реализира чрез *генеричен „web-scraping“*, последван от *машинно самообучителен* подход. Резултатът е списък с характеристики на социалните медии, като за сравнителен анализ се използва статистическото изследване „Използване на ИКТ от предприятията“.

Четвъртият „use-case“ „Генериране на списък с URL адреси на предприятията“ чрез прилагане на италиански софтуер (разработен от ISTAT) е аналогичен на първия чрез използване на същите методи.

Не по-малко важен е въпросът за **програмните езици, ИТ инструментите и библиотеките,** използвани в пилотните „use-cases“. ИТ средствата за извършване на „web-scraping“, както вече беше отбелязано, биват генерични и специфични. Типичните генерични инструменти са следните: „import.io“, „Scrapy“, „imacros“, „Apache Nutch“ и други подобни. Втората група инструменти включва библиотеки за конкретни цели, като програмите „Tweepy за Python“, използвани за извличане на данните от Twitter. Подробен списък с web-scraping инструменти и библиотеки е достъпен в различни хранилища, включително в Github. Благодарение на разнообразието от инструменти и методи за „web-scraping“ всеки пилотен проект можеше да бъде реализиран по различен начин. Една от целите на текущото емпирично изследване от технологична гледна точка беше да се избере популярен, с отворен код или безплатен софтуер, който служителите на НСИ познават. Това е основната причина да се използва традиционен, както и специфичен софтуер за Big Data, който може да бъде изтеглен и приложен в НСИ без допълнителни разходи.

Целта на избора на инструменти за средства за съхранение е да се създаде среда, която да се поддържа лесно и CSV файловете да са с възможност за общ избор. Най-разпространеният начин за съхраняване на данни е файлова система от типа CSV файл. Дефинираните пилоти използват MySQL (релационни бази данни) и NoSQL Apache Solr (нерелационни бази данни). Решението за използване на файловата система като първично съхранение на данни е резултат от факта, че повечето от инструментите, използвани в пилотите, имат вградени библиотеки за достъп до CSV файлове. От друга страна, CSV файловете се използват за съхраняване на резултатите от анализа. Използването

на този тип файл позволява зареждането на данните в популярни приложения като R или MS Excel. За да се увеличи производителността на достъп до такива файлове, има възможност те да се съхраняват в HDFS файлова система (Hadoop Distributed File System), за да се извърши автоматичен и много ефективен паралелен достъп до данните. Целта на използването на Apache Solr в „use-case“ № 4 е да се осигури мащабируема среда, която да може да съхранява различни видове данни. Въпреки това основната цел на използването на Apache Solr е да съхранява уебсайтовете на предприятията в NoSQL база данни. Този тип база данни позволява динамично търсене чрез нейното съхранение, включително пълно текстово търсене, подчертаване на удари, фасетирано търсене, динамично клъстеризиране, интегриране на базата данни, обработка на богати документи, разпределено търсене, репликация на индекси и висока скалируемост.

За целите на настоящото емпирично изследване бяха разработени няколко различни софтуерни инструмента. Например ISTAT е разработил софтуер за URLSearcher в Java, който позволява да се извличат URL адреси за различни предприятия въз основа на техните атрибути, като име, град, адрес за връзка. Този софтуер е приложим за изпълнението на „use-case“ № 4 върху набор от данни на НСИ, тъй като може да се използва в Java среда във всяка операционна система. Резултатът е CSV файл или набор от файлове, включително идентифицирани връзки. По-нататъшен анализ на URL адресите се извършва с други потребителски Java приложения за търсене на URL адреси: URLSearcher, RootJuice и UrlScorer. Софтуерът е разработен като кратки скриптове, използвани за конкретни действия относно събирането и обработката на данни. Най-популярната среда за тестване на пилотите е използването на виртуална машина (например Linux Mint) или специална машина (например Linux Ubuntu Server, MS Windows).

Придобитият опит по време на изпълнението на четирите пилотни проекта показва, че софтуерните скриптове и приложения са независими от ИТ платформата. Например възможно е да се стартират Python скриптове или Java приложения в Linux, както и в Windows среда без никакви промени в изходния код. Ефективността на изпълнението е ключов проблем, особено когато се изтеглят и обработват целите уебсайтове. Обработката на неструктурирана информация изисква голям обем на централния процесор и памет, особено при прилагане на алгоритми за машинно обучение и като резултат обработката на данните не е много ефективна. Поради факта, че по-голямата част от средата, използвана за „сценариите“, има ограничени ресурси на процесора и

паметта, е трудно да се прецени колко ефикасен ще бъде алгоритъмът в реална производствена среда.

Въз основа на обработката на данните на НСИ със специално разработения за целта софтуер може да се направи заключение, че конвенционалните ИТ инструменти са достатъчни за създаването на списък с URL адреси за уебсайтовете на няколко десетки хиляди предприятия. Софтуерът, разработен и използван за внедряването на пилотите, е безплатен и с отворен код макар някои негови компоненти да са разработени „по поръчка“ за специфични цели. Това означава, че всеки може лесно да тества и подобри всеки от инструментите. От друга страна, при използването на такива софтуерни средства невинаги можем да разчитаме на добра документация или на подробно ръководство, за да се направи всичко работещо.

Въз основа на прилагането на италианския софтуер за масива от данни на НСИ може да се твърди, че Apache Solr има технически проблеми по отношение на Solr Connection pool в етапа на зареждане, веднага след фазата на „извличане“ на данните. Обстоятелството, че ИТ средствата за обработка на Big Data се променят много често, а паралелно с това - и технологията на уебсайтовете, е необходимо да се осигури възможно най-гъвкаво разработване на този тип ИТ инструменти. Това поставя въпроса за устойчивостта на използваните технологии. ИТ инструментите за пилотите вероятно ще се променят през следващите няколко години и затова не се препоръчва използване на конкретен компютърен език, за да е възможно лесно да се премине към други платформи.

Въз основа на българския опит изискванията за съхранението на база данни за около 27 000 предприятия отнемат памет около 1 GB твърд диск, включително данни от бизнес регистъра, извлечени данни чрез търсене на API, уебсайтове на предприятията и електронни магазини от заглавия на първа страница, ключови думи, описания и данни от URL адреси. Това позволява да се направи заключение, че за тези конкретни „сценарии“ е възможно да се използват традиционни технологии. Както вече беше споменато, възможен е избор между файлова система (CSV, JSON и т.н.), NoSQL база данни (Solr, Cassandra, Hbase и т.н.) или релационна база данни (MySQL, PostgreSQL, SQL Server и други). Решението за използването на конкретно средство за съхранение на информацията трябва да се вземе в зависимост от обема и вида на данните, които ще бъдат съхранявани. В този контекст е и въпросът с дублирането на данните. Необходима е рамка за премахване на дублирането, която автоматично ще изключи всички дублирания на уебсайтове и конкретна информация, взета от тях.

Като възможен резултат от разработените пилотни проекти са изчислени някои допълнителни индикатори като: относителен дял на извлечените URL адреси от списъка на предприятията, относителен дял на предприятията, ангажирани с електронната търговия чрез уебсайтовете на предприятията, относителен дял на предприятията, които присъстват в социалните медии на уебсайтовете на предприятията, брой заети и други показатели на ниво област. За съжаление, все още тези резултати са достъпни само чрез Интранет мрежата на НСИ и не са публично достъпни. Прогнозните стойности, получени от източниците на Big Data, могат да се използват за постигане на две цели: на ниво единица - за обогатяване на информацията, съдържаща се в регистъра на изучаваната съвкупност, и за получаване на оценки на ниво съвкупност. Качеството на данните на ниво единица може да бъде измерено, като се вземат предвид същите показатели, произведени за оценка на модела, пасващ на обучителния набор от данни. При определени условия (ако обучителния набор от данни е представителен за цялата съвкупност) измерването на точността, както и на чувствителността и специфичността, изчислени в тази подгрупа, може да се счита за добра оценка на качеството на получените резултати от събраните Big Data.

Обратно, измерването на качеството на оценките на съвкупността, за които се използват прогнозни стойности, е много по-сложна. Пълното измерване трябва да се основава на оценка на средната квадратична грешка, т.е. съвместно отчитане на вариацията и изместването, влияещи на оценките на параметрите на съвкупността. Съществуват случаи, при които е възможно изчисляването на вариацията на оценителя, като се прилагат методи за повторно проектиране на извадката (по-специално bootstrap), вместо да се оценява изместването, изискващо познаване на истинската стойност на параметъра в съвкупността, което е рядко срещано условие на практика.
Симулационните проучвания, при които се генерира изкуствена съвкупност, споделяща разпределителни характеристики с истинската съвкупност, могат да подпомагат оценяването на всички компоненти на средната грешка. Например в „сценариите“ за електронната търговия и представянето в социалните медии относителният дял на предприятията, чиито уебсайтове се характеризират с положителен отговор на тези променливи, може да бъде оценен в различни области чрез данни от традиционни статистически изследвания и чрез използване на интернет данни (с подход, базиран на модели).

Систематичната грешка при подбора на единиците, произтичаща от идентификацията на уебсайта, може да бъде особено важна - методите за идентифициране на URL адреси са по-склонни да работят добре за уебсайтове, които извършват електронна търговия. Двата набора от оценки могат

да бъдат сравнени. За да се реши дали тяхната разлика по отношение на качеството е релевантна, сравнението може да използва техники за повторно проектиране на извадката и симулационни проучвания.

II. Практическа реализация на пилотните „сценарии“

1. Use-case 1: Генериране на списък с URL адреси на предприятията (URLs retrieval)

В пилотния проект е използвана информация от статистическия Бизнес регистър на НСИ, предприятията от който формират съвкупността на изследването „Използване на ИКТ от предприятията“. Основната цел се състои в генериране на списък от валидирани URL адреси на предприятията, който да бъде използван впоследствие за извличане на информация от корпоративните сайтове за наличие на електронна търговия и присъствие в социалните медии („use-case 2“ и „use-case 3“) чрез прилагане на различни ИТ техники. Екипът за реализацията на този „сценарий“ включва: ИТ специалист/програмист - да извършва дейностите по извличане на URL адреси от наличните набори с данни и да разработи софтуерни програми за прилагане на техниката „web-scraping; експерт, отговорен за поддържане на актуален списък с URL с адреси на предприятията; статистик, за да анализира концептуалното свързване между уебинформацията и бизнес регистъра, като решава възникнали методологични проблеми - например липсващи стойности на уебсайта на предприятието и др., и технически персонал. Работният процес за настоящия „use-case“ се състои от следните последователни действия: създаване на тестови набор от предприятия със свързани URL адреси от посочената целева съвкупност; използване на приложен програмен интерфейс API в мрежата за търсене името на предприятието или името на предприятието, последвано от „контакт“, както и съхраняване на първите 10 резултата като „кандидати“-уебсайтове.

За целите на настоящия „сценарий“ е разработен специален софтуер, като се използва определен брой API търсения от интернет търсачките - *Jabse* и *Google Search.* За всеки „кандидат“-уебсайт се оползотворяват събраните уебданни, за да се получи подробна информация. Това могат да бъдат извлечени данни от уебсайта на компанията или фрагмент от резултата от API за търсене. Използването на събраните данни за идентифициране на уебсайтове се извършва ръчно или чрез прилагане на автоматичен алгоритъм. В процеса на работа по този пилотен проект се достигна до следните няколко извода:

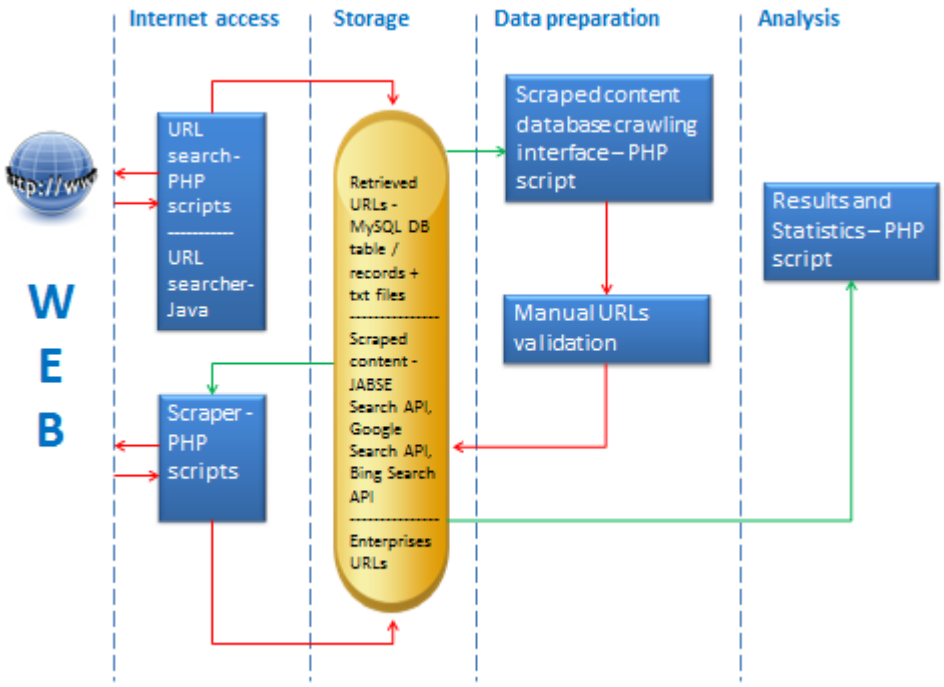
- Чрез прилагането на основната методология за изпълняване на заявки за търсене (API Search) на уебсайта на предприятието е възможно да се идентифицират голям брой уебсайтове на

същото предприятие и след това да се осъществи свързване на уебданните с неговите характеристики.

- Възможно е да се идентифицират както фалшиви положителни данни (неправилно идентифицирани уебсайтове), така и фалшиви негативни данни (уебсайтове, които не са идентифицирани). За проверка на „гранични“ случаи се изисква ръчна проверка, извършена от статистически експерти - особено при значително малък набор от данни. За да се създаде списък на URL адреси на предприятията с добра точност, винаги ще е необходима известна „експертна“ помощ.
- Разнообразието от уебданни може да е полезно при избора между „кандидат“ уебсайтове API Search - включително данни, извлечени директно от уебсайтовете, „фрагменти“ от резултатите от търсенето, класиране на резултатите от търсенето и информация за регистрация на уебсайтове. Обаче трябва да се има предвид, че не е достатъчно да се използва само регистрационна информация на предприятието, което предполага, че някои данни трябва да бъдат събрани от уебсайтове. Това означава, че всяка официална статистическа институция, която се интересува от идентифициране на уебсайтове на предприятия, трябва предварително да е разрешила всички правни и етични въпроси, произтичащи от националното законодателство по отношение на прилагането на техниката „web-scraping“.
- Когато методът за идентифициране на уебсайтове не може да идентифицира всички уебсайтове на предприятието, е важно да се отчете стеминг отклонението от вероятността да бъде идентифициран всеки от дадените уебсайтове.

Нека си припомним, че целевата съвкупност, обект на „web-scraping“, се състои от 26 836 предприятия, 20 649 имейла и 2 006 URL адреса. Използваните подходи са основно два: приложение на Jabse Search API и Google Custom Search API на базата на името на предприятието с филтриране на резултатите от Search API и извличане на URL адреси от набори от данни, съдържащи Булстат номер, URL адреси, данни за контакт и други характеристики на предприятието от статистическия Бизнес регистър. Обща представа за производствения поток може да се получи от фиг. 2, на която е представена логическата архитектура на Use-case 1.

Фиг. 2. Логическа архитектура на Use-case 1: URLs retrieval



Компонентът „URL Searcher“ използва информация за URL адреси и имейли от бизнес регистъра, за да проверява, верефицира и генерира имена на домейни и да обработи списъка с URL адреси. Средството Scraper използва фирмените имена и чрез ефективната работа на интернет търсачките Jabse Search API и Google Custom Search API се събира набор от 10 предложени/вероятни URL адреси на дадено предприятие от съвкупността. Събраната информация от URL Searcher и Scraper се съхранява в специално създадена за целта база данни. Истинските URL адреси на предприятията се идентифицират чрез софтуерен интерфейс (web-crawling), специално разработен за „обхождане“ на базата данни, които статистическите експерти на последната фаза от процеса валидират ръчно. Във фазата на анализ статистическите резултати са изчислени със специфичен софтуерен скрипт. След като компонентът URL Searcher получи списък с предприятия с налични URL адреси и имейли, същият проверява дали първоначалните URL адреси са реални уебсайтове и съхранява резултатите в базата данни. Ако URL адресите не са потвърдени или липсват, то тогава URL адресите се генерират от имената на домейните на електронната поща, като се изключват популярни домейн наименования на имейлите (като gmail, yahoo и др.), в случай че са налични. Генерираните URL адреси се верифицират за действителното им съществуване от Searcher. Всички

проверени и установени като действителни URL адреси са съхранени в базата данни (**7 038 URL адреса**).

Компонентът Scraper използва автоматизирания интерфейс за търсене на *Jabse*, като получава до 10 резултата от търсенето за наименованието на дадено предприятие на български език и до 10 резултата от търсенето за наименованието на предприятието, транслирано на латиница. Тогава Scraper изключва от резултатите от търсенето сложните URL адреси и записва само тези, които са близки до наименованието на домейна, като предполага, че това са най-вероятните адреси. Резултатите се **съхраняват в базата данни в текстов и html формат (15 638 серии** с до 10 най-вероятни резултата от търсенето на български език, **16 201 серии** с до 10 най-вероятни резултата от търсенето на латиница). След това Scraper използва интерфейса за търсене на Google, като получава до 10 резултата от търсенето за наименованието на фирмата на български език и записва резултатите в базата данни в **json формат** (**26 829 серии** с до 10 резултата от търсенето). Интерфейсът за „обхождане“ на базата данни е използван като помощно средство от статистическите експерти, за да избират реалните URL адреси на фирмите от всички предложени/намерени URL адреси от URL Searcher и Scraper. Резултатите от тази фаза са **9 809 реални и действителни URL адреса.** Скриптът за резултати и статистики дава информация за корпоративните URL адреси в реално време.

Що се отнася до технологичния избор към момента на изпълнение на този пилотен проект („use-case 1“) НСИ нямаше никакъв опит в областта Big Data и по отношение на прилагането на техниката „web-scrapping“. Поради тази причина първият избор на технологични инструменти са безплатни софтуерни уебприложения, с които само един експерт в НСИ имаше известен опит: Apache уебсървър, MySQL база данни и PHP програмен език. В процеса на работа са използвани PHP за програмния софтуер, MySQL за платформата за съхранение и Apache за изпълнение на PHP скриптове през уеббраузърите. В специално разработения по поръчка софтуер са интегрирани и използвани предложените/намерените резултати от уебприложенията *Jabse Search API* и *Google Custom Search API).* PHP скриптовете бяха изпълнени в браузър с използване на метамаркер за опресняване на HTML съдържание (например скриптът задава заявки за приложния програмен интерфейс (Search API) на всеки три секунди с фирмени данни и съхранява информацията, събирайки я в базата данни.

След успешното завършване на практическата реализация, могат да се направят няколко **заключения:**

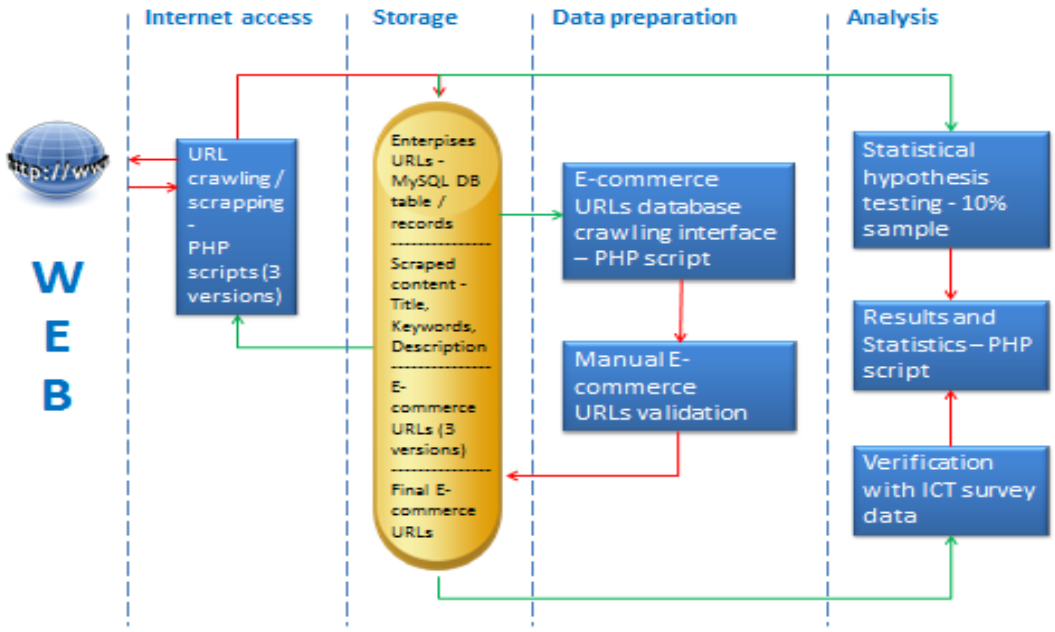
- По отношение на методологията може да се твърди, че *Google Search API* дава най-добри резултати, като осигурява 200 търсения на ден безплатно или 1 000 търсения за сумата от 5 EUR до максимум 10 000 търсения на ден. От друга страна, базата данни на Jabse не обхваща всички предприятия. Jabse Search работи по-добре с английската си версия отколкото с българската, но за съжаление Search API покрива само българската версия. Като цяло **26 836 записа бяха проверени ръчно от експерти (част от екипа на проекта) в рамките на 45 работни дни или средно проверени по 600 записа за всеки работен ден.**
- Конвенционалните ИТ инструменти са достатъчни за създаването на списък с URL адреси на десетки хиляди предприятия. Размерът на базата данни, съдържаща около 27 000 предприятия, заема около 1 GB хардуерно пространство (данни от Бизнес регистъра, данни от търсачките Search API, уебсайтове на предприятия, електронни магазини от наименованията на първите страници на уебсайта, ключови думи, описания и URL адреси).
- Няма правни ограничения, защото са използвани публично достъпни ИТ инструменти и продукти (Jabse и Google Search APIs), за да се получат URL адресите на предприятията.

2. Use-case 2: Електронна търговия в предприятията (E-commerce)

За изпълнение на задачите по настоящия пилотен проект е използван списъкът с действителните URL адреси на предприятията (получени като резултат от работата по „use-case 1“), за да се установи дали дадено предприятие, което има фирмен уебсайт, се занимава с електронна търговия. Основната цел на този „use-case“ е прилагане на техниката „web-scraping“ върху заглавните уебстраници на официалните уебсайтове на фирмите и след това да може да се направи възможно най-точна прогноза за наличието на е-търговия на корпоративния сайт. Това става чрез използване на предварително дефинирана таксономия и набор от ключови думи, приложени към текстовото съдържание на уебсайта, получено чрез извличане на съдържание от текст. За да се постигнат по-качествени резултати, извличането на съдържание от текст е направено по три различни начина (три прогнозни версии). Получените крайни прогнозни резултати бяха валидирани ръчно, за да се оцени коя от трите прогнозни версии е най-точна. Очакваните резултати са дефинирани алтернативно: „успех“ - положителна или отрицателна оценка (дадено предприятие извършва или не извършва електронна търговия), и „неуспех“ - няма достъп до уебстраницата на предприятието. Работният процес за настоящия „use-case“ се състои от следните последователни действия: извличане на текстово съдържание от списъка с предварително дефинирани интернет страници на предприятията (от „use-case 1“) само на ниво заглавна страница на уебсайта на дадено

предприятие. Обемът на извадката, от която се извлича текстово съдържание, е 9 809 уебсайта на предприятия; създаване на характеристики въз основа на наличието или отсъствието на думи в текстовото съдържание: създадени са списъци или „речници“ на ключови думи (няколко групи от подобни списъци), които индикират наличието или отсъствието на електронна търговия на сайта на дадено предприятие; използване на тези характеристики и някои алгоритми, за да се предскаже дали дадено предприятие е ангажирано в електронната търговия, а именно прилагане на автоматичен филтър, базиран на създадените характеристики в предишния етап на работа. Като цяло опитът за идентифициране на сайтовете за електронна търговия се оказа успешен, но усилията бяха насочени към постигане на баланс между прецизност и повторяемост на изпълнението на задачата. Поради тази причина възникна необходимостта от по-нататъшно разработване на специфични методи, преди да се стигне до получаване на надеждни оценки. Изпълнението на един по-опростен метод, основан на правила за идентифициране на създадените характеристики, може да се оцени като сравнително добро. Въпреки това ефективното прилагане на модел от типа „кошница с думи“, където всяка дума се третира независимо и използването на по-съвременни техники - като например от типа NLP техники, биха дали вероятно още по-добри резултати. Логическата архитектура на „use-case 2“ е представена на фиг. 3.

Фиг. 3. Логическа архитектура на Use-case 2: E-commerce



Средството *URL crawling-scrapping* използва информация от списъка с URL адресите, за да „посети“ уебстраниците на предприятията, да предскаже URL адресите за електронната търговия на предприятията и да извлече заглавие, ключови думи и описателно съдържание от първата страница на фирмения уебсайт или да предскаже първите страници на уебсайтове за електронна търговия. Събраната по този начин информация е съхранена в базата данни. Интерфейсът за „обхождане“ на базата данни (*URLs database crawling interface*) за електронна търговия се прилага като помощно средство от статистически експерти-доброволци. Същите проверяват и валидират ръчно коректния URL адрес за електронна търговия на предприятието, използвайки информацията, събрана от средството *URL crawling-scrapping*. Във фазата на анализа бяха приложени класическите методи за проверка на статистически хипотези, за да се провери колко точни и изчерпателни са използваните „web-scraping“ алгоритми за предсказване. В допълнение, окончателните резултати от този пилотен проект бяха сравнени и верифицирани с данните от традиционното статистическо изследване „Използване на ИКТ в педприятията“ за 2016 г., провеждано от НСИ. Статистическите резултати от сравнението са изчислени със специфичен софтуерен скрипт. Средството *URL crawling-scrapping* взема уебадреса от списъка с URL адресите на предприятията и извлича съдържанието на първата страница на фирмения уебсайт. След това с помощта на PHP скрипт с три логически алгоритъма

(използвайки 4 положителни и 1 негативен списък с ключови думи) прогнозира URL за електронната търговия на предприятието, ако тя съществува. Резултатите са, както следва: чрез алгоритъм № 1 са прогнозирани 1 139 URL адреса за е-търговия; чрез алгоритъм № 2 - 1 048 URL адреса, и чрез алгоритъм № 3 - 662 URL адреса за електронна търговия на предприятията. *URL crawling-scrapping* компонента „извлича“ уебадреса на фирмата за електронна търговия, заглавието, ключовите думи и описанието на уебсайта и съхранява извлечената информация в SQL база данни заедно с прогнозираните вероятни URL адреси за електронна търговия. Интерфейсът за обхождане на базата данни за електронната търговия с URL адреси (*e-commerce URLs database crawling interface*) представя резултати от предвиждането за е-търговия, първите уебстраници на електронните магазини и заглавията на първите страници на предприятията, ключови думи и текстови описания. На следващ етап статистически експерти-доброволци валидираха получените резултати ръчно и само реалните URL адреси за е-търговия на предприятията бяха съхранени окончателно в базата данни. По този начин са открити общо **856 уебстраници за електронна търговия.**

Отрицателната оценка е тествана с прилагане на класическия метод за доказване на статистически хипотези (10% извадка) при допускане за 90% точност и 80% пълнота на прилагания алгоритъм. За целта беше излъчена 10% случайна извадка от двете подсъвкупности: е-търговци и останалите, които не са такива, като статистически се проверява дали точността е по-ниска от 90% и пълнотата е по-малка от 80%. Прилагайки нормалното разпределение за нулевата хипотеза, се стига до извода, че използваният филтър е точен и изчерпателен. В резултат на това беше установено, че едва 27 е-търговци не са обхванати от алгоритмите за прогнозиране. В допълнение, очакваният резултат от този пилотен проект беше да се вземе решение дали произведената информация по този иновативен начин може да се използва за замяна на някои въпроси, съдържащи се във въпросника на традиционното изследване „Използване на ИКТ от предприятията“, с цел намаляване на тежестта на респондентите и получаване на реална, но по-качествена информация. Верификацията на резултатите е извършена с данните от изследването „Използване на ИКТ от предприятията“. След бенчмаркинг анализа между данните от традиционното изследване в областта на ИКТ и получената от настоящия пилотен проект информация, се получават следните резултати: от 26 836 предприятия (обхват на проекта) в извадката за 2016 г. на изследването попадат 4 332 предприятия от тях. Намерени са 89 нови предприятия, извършващи в действителност електронна търговия, които са включени в обхвата на традиционното изследване, но дават отрицателни отговори при попълване на въпросника от анкетното проучване. Като правило, ако е-магазините на предприятията имат основни

уебсайтове или имат връзка към електронни магазини, то те в повечето случаи се откриват на първите страници от уебсайтовете на предприятията. В някои случаи специалистът, разработващ фирмен уебсайт, поставя връзка към собствения си електронен магазин, така че трябва да се използва негативен списък от ключови думи със създателите на уебсайтове. Трябва да се има предвид, че по-дългият списък с ключови думи невинаги дава по-добри резултати за разлика от варианта, когато по-стриктният алгоритъм дава по-точни резултати, но пък пропуска повече URL адреси за е-търговия.

След успешното завършване на практическата реализация, могат да се направят няколко **заключения:**

- По-свободният алгоритъм намира повече уебсайтове, които в действителност не са електронни магазини. Ако от излъчената 10% извадка, са получени 856 уебстраници за електронна търговия и са открити 27 пропуснати е-търговци, то тогава се получават общо 1 126 вероятни URL адреса на фирми за е-търговия (856 + 27 x 10 = 1126). Резултатът е близък до прогнозирания в началото брой от 1 139 URL адреса за е-търговия на предприятията, прилагайки по-свободния алгоритъм. Като заключение, общо 11.5% от предприятията, които имат уебсайтове, извършват е-търговия и 4.2% от общия брой предприятия в съвкупността извършват електронна търговия.
- Използваните ИТ средства и инструменти са достатъчно ефективни за изпълнението на настоящия пилотен проект.
- Не са възникнали правни ограничения по отношение на прилагането на техниката „web-scrapping“.

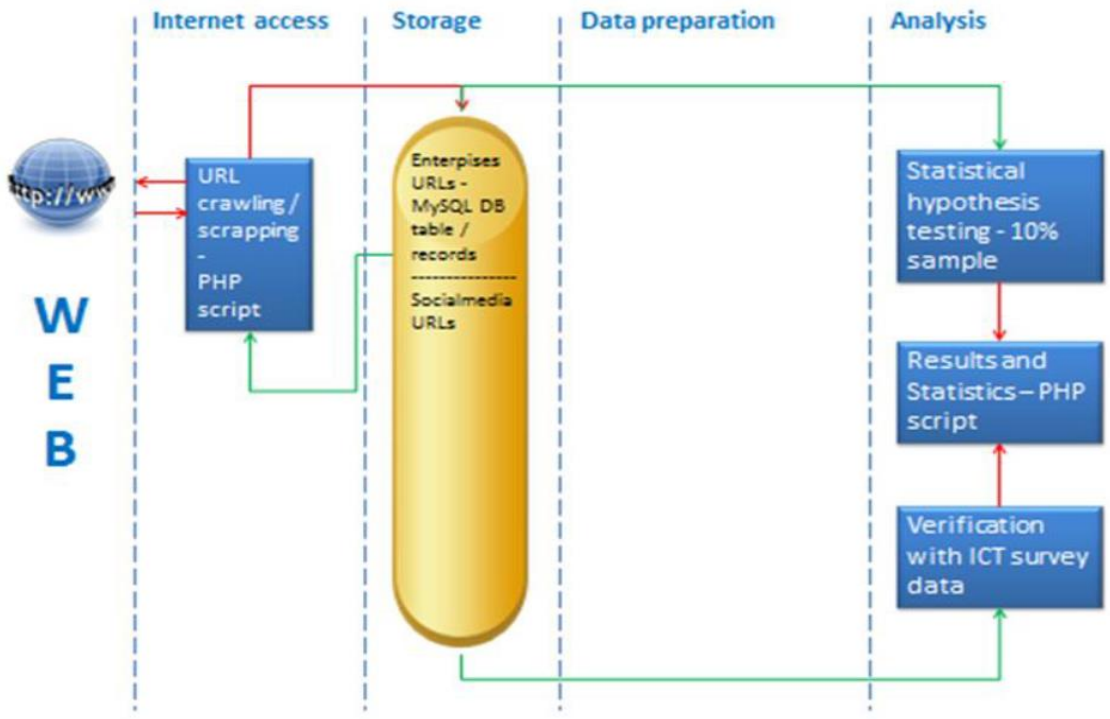
3. Use-case 3: Присъствие на предприятията в социални медии (Social media presence)

За изпълнение на дейностите се използва списъкът с URL адресите на предприятията (получен като резултат от изпълнението на „use-case 1“), за да се провери дали дадено предприятие присъства, или не в социалните медии (Facebook, Twitter, Linkedin, Google, You tube, Pinterest, Instagram). Основната цел на този пилотен проект е да предостави информация за активността на българските фирми в социалните медии, т.е. всички профили в социалните медии са взети под внимание. Общата концепция за този „use-case“ е да се извлече уебстраницата и после да се търсят връзки към профилите в социалните медии, включени в обхвата на този проект. В бъдеще е възможно да се добавят още атрибути за събиране на информация, например дали профилът е актуален и колко често се променя съдържанието му. Очакваният резултат от този пилот е вземане

на решение дали данните за социалната медийна активност на предприятията могат да се използват за актуализиране на статистическия бизнес регистър и дали има засилено присъствие на българските компании в социалните медии. В допълнение, очакваните резултати могат да бъдат използвани за вземане на решение за замяна на някои въпроси, съдържащи се във въпросника на изследването „Използване на ИКТ от предприятията“. Работният процес за настоящия „use-case“ се състои от следните последователни действия: извличане на съдържанието на първите уебстраници с URL адреси от списъка с URL адреси на предприятията (получен като резултат от изпълнението на „use-case 1“); идентифициране на профилите на социални медии (предимно Facebook и Twitter) чрез проверка на връзките от извлечените страници и филтриране с наименованията на социалните медии; оценка на резултата чрез бенчмарк анализ с данните от традиционното статистическо изследване „Използване на ИКТ от предприятията“. Изводът, до който достигна екипът по време на работата по този „сценарий“, е: в повечето случаи профилите на предприятията в социалните медии са успешно идентифицирани. Въпреки това намерените профили чрез прилагане на „web-scraping“ са по-малко като обем от тези, получени като отговори на въпросника от изследването „Използване на ИКТ от предприятията“. Често срещан проблем е още, че идентифицираните профили в социалните медии не са непременно съответстващи на търсените URL адреси или предприятия. Не може да се твърди със сигурност дали фирмените профили в различните социални медии се актуализират, или не. Анализът на данните за социалните медии с Twitter API е ефективен, но възникват трудности с текущия Facebook API.

Екипът на проекта използва идентичен набор от ИТ средства за този пилотен проект, както за „use-case 1“ и „use-case 2“, а именно: ***специално разработен софтуерен скрипт*** за „извличане“ на уебсъдържанието от сайтовете на предприятията и подпомагане на експертите да определят присъствието на дадено предприятие в социалните мрежи. Производственият поток на „use-case 3“ е представен на фиг. 4.

Фиг. 4. Логическа архитектура на Use-case 3: Social media presence



Средството „URL crawling-scrapping“ (същият инструмент е използван в „use-case 2“) използва информация от списъка с URL адресите на предприятията, за да „посети“ уебстраниците на предприятията и да прогнозира присъствието на предприятията в социалните медии. Информацията, събрана от този компонент, е съхранена в базата данни. Окончателните резултати са сравнени с данните от изследването „Използване на ИКТ от предприятията“ за 2016 година. Статистическите резултати са изчислени със специфичен софтуерен скрипт. Функционалното описание на компонента е следното: „URL crawling-scrapping“ „взема“ уебадреса от списъка с URL адреси на предприятията (9 809 URL адреса от „use-case 1“) и извлича съдържанието на първата страница на уебсайта на предприятието. След това, използвайки името на социалните медии, PHP скриптът търси уеблинк към профила на социалните медии и съхранява намерената информация в SQL база данни. Получените резултати са, както следва: Facebook - 2 356 профила; Twitter - 922 профила; Linkedin - 560 профила; Google - 871 профила; Youtube - 527 профила; Pinterest - 139 профила, и Instagram - 127 профила.

Установено е, че 24.9% от предприятията с уебсайтове имат поне един профил в социалните медии, а 9.1% от предприятията използват поне една от обхванатите в изследването социални медии.

Получените резултати от „web-scraping“ са проверени за прецизност и изчерпателност чрез прилагане на методите за проверка на статистически хипотези. За целта беше излъчена 10% случайна извадка при гаранционна вероятност за 90% точност и 80% пълнота на прилагания алгоритъм. По-конкретно, беше тествано дали точността е по-ниска от 90% и пълнотата е по-малка от 80%. Прилагайки нормалното разпределение за нулевата хипотеза, се стига до извода, че използваният филтър е точен и изчерпателен. В резултат на това беше установено, **че само четири предприятия не са обхванати.** Верификацията на резултатите е извършена с данните от изследването „Използване на ИКТ от предприятията“. След бенчмаркинг анализа между данните от традиционното изследване в областта на ИКТ и получената от настоящия пилотен проект информация, се получават следните резултати: от 26 836 предприятия (обхват на проекта) в извадката за 2016 г. на изследването попадат 4 332 предприятия от тях. Намерени са 382 нови предприятия, които присъстват в социалните медии и попадат в обхвата на изследването.

След успешното завършване на практическата реализация, могат да се направят няколко **заключения:**

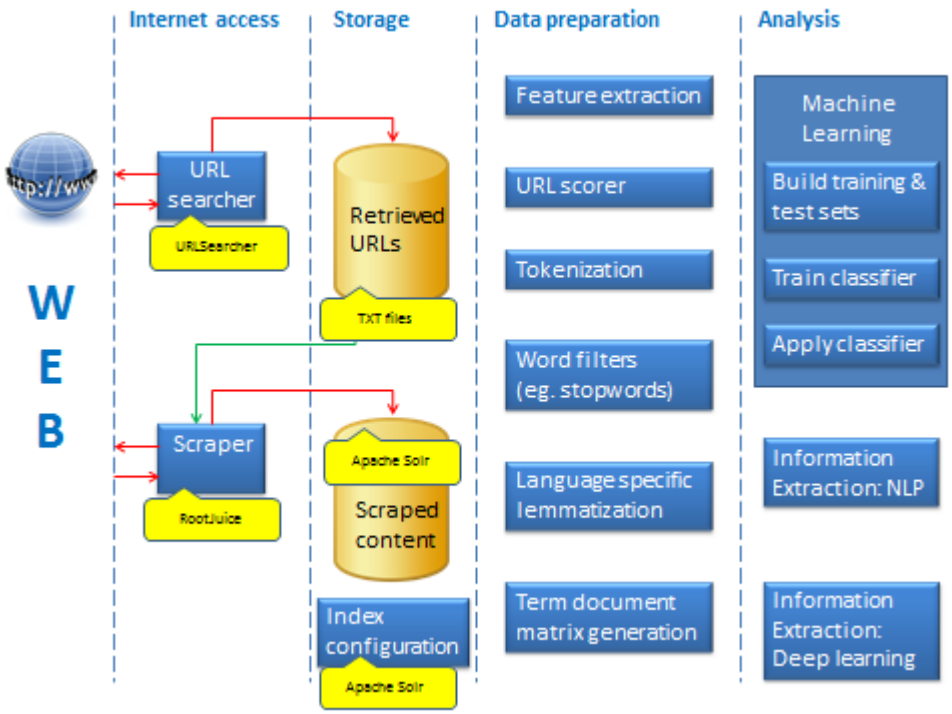
- По отношение на методологията съществува риск, че някои от връзките във Facebook или Twitter, представени на уебстраницата, може да са свързани с други предприятия. Поради тази причина е необходима оценка на Facebook и Twitter профилите, за да се предостави надеждна информация. В някои случаи предприятията могат да имат няколко Facebook профила. Поради тази причина е необходимо основният социален профил да се свърже с профила на компанията. Като правило връзките на профилите на социалните медии с профилите на предприятията се намират на първите страници на уебсайтовете на предприятията.
- По отношение на ИТ технологии е направен изводът, че използваните ИТ средства и инструменти са достатъчно ефективни за изпълнението на настоящия пилотен проект.
- И при този пилотен проект не са възникнали правни ограничения по отношение на прилагането на техниките „web-търсене“ и „web-scrapping“.

4. Use-case 4: Генериране на списък с фирмени URL адреси на предприятията чрез прилагане на италианския софтуер (разработен от ISTAT)

Целта на пилотния проект е тестване на софтуера на ISTAT за извличане на данни и инвентаризация на URL адреси с автоматизирана процедура за извличане на URL адреси в български условия. Базовата съвкупност е същата - предприятията в обхвата на изследването „Използване на ИКТ от предприятията“. Задачите са аналогични с тези на „use-case 1“ - генериране на списък от

URL адреси на предприятията, които да бъдат използвани впоследствие за „извличане“ на информация за е-търговия и присъствие в социалните мрежи. НСИ използва предложения софтуер с отворен код от ISTAT: Java Run time environment за програмите URLSearcher, RootJuice, URLScorer и URLMatchTableGenerator и Apache Solr платформа за съхранение. Решението за тестването на италианския софтуер върху българската съвкупност от предприятия беше взето с цел да се споделят добри практики и опит между европейските страни, участващи в проекта. Резултатите от този „use-case“ са използвани за бенчмаркинг с резултатите от „use-case 1“, като са направени оценки за постигнатите резултати в двата случая. Работният процес за настоящия „use-case“ се състои от следните последователни действия: използване на тестовия списък от предприятия със свързани URL адреси от „use-case 1“; използване на приложен програмен интерфейс (API) за търсене в мрежата на името на предприятието или името на предприятието, последвано от „контакт“ и съхраняване на първите 10 резултата като „кандидати“-уебсайтове. За целите на настоящия „сценарий“ е приложен италиански софтуер, като се използва определен брой API търсения от интернет търсачката Bing; за всеки „кандидат“-уебсайт се оползотворяват събраните уебданни, за да се получи подробна информация. Това могат да бъдат извлечени данни от уебсайта на компанията или фрагмент от резултата от API за търсене и използване на събраните данни за идентифициране на уебсайтове чрез прилагане на автоматичен алгоритъм или ръчно. Технологичната реализация на проекта е представена на фиг. 5.

Фиг. 5. Логическа архитектура на Use-case 4: Генериране на списък с фирмени URL адреси на предприятията чрез прилагане на италианския софтуер



Накратко, общото описание на производствения поток в логическата архитектура може да се представи по следния начин: средството URLSearcher използва фирмените имена с приложния програмен интерфейс API на Bing Search, за да може да се формира набор от 10 на брой предложени URL адреса и събраната информация се съхранява в txt файл. Компонентът RootJuice „взема“ txt файла, „извлича“ съдържанието на фирмените уебсайтове и записва информацията в csv файл. Информацията от csv файла се зарежда в Apache Solr (платформа за търсене на предприятия с отворен код), а UrlScorer използва съхранените вече данни в Apache Solr, за да генерира файл със зададени резултати за всеки от предложените URL адреси на предприятието. След това компонентът URLMatchTableGenerator получава резултатите от URLScorer и ги сравнява с предварително известен списък на URL адресите на предприятието. Получените резултати се анализират и се избира най-вероятният уебадрес на дадено предприятие. Във функционално отношение компонентът URLSearcher получава два файла: единият, съдържащ списъка с фирмените имена, и съответният списък с идентификационни номера на предприятията от бизнес регистъра (както отбелязахме 26 836 предприятия). За всяко предприятие Bing търсачката „извлича“ списък с първите 10 URL

адреса, които се съхраняват във файл (всяка фирма има по един такъв файл). В края на тази дейност програмата чете всеки произведен файл и създава сийд файл във формат txt, съдържащ всички резултати. Програмата RootJuice приема като вход 3 файла: сийд файл от URLSearcher, списък с домейни на URL адреси, които трябва да се избягват (обикновено домейни на директории, жълти страници и т.н.) и конфигуриращ файл. Програмата RootJuice се опитва да достигне до HTML страница за всеки ред от сийд файла (ако URL адресът не е в списъка на домейните, които трябва да се избягват). От всяка достигната HTML страница програмата избира само текстовото съдържание на полетата, които представляват интерес, и ги записва като отделен ред CSV файл. След това CSV файлът от RootJuice се импортира в платформата за съхранение с отворен код Apache Solr. UrlScorer е програма, която чете един по един всички документи, съдържащи се в определена Solr колекция, и определя оценки на всеки от тези документи въз основа на стойностите на някои показатели. По-специално, изчислява се стойността на двоичните показатели, например: URL адресът съдържа деноминацията (Да/Не); „извлеченият“ уебсайт съдържа географска информация, съвпадаща с вече наличната в бизнес регистъра (Да/Не); „извлеченият“ уебсайт съдържа същия фискален код в бизнес регистъра (Да/Не); „извлеченият“ уебсайт съдържа същия телефонен номер в регистъра (Да/Не) и т.н. URL Match TableGenerator получава резултатите от URLScorer и ги сравнява с предварително известен списък от URL адреси на предприятието. Резултатът показва, че софтуерът прогнозира правилните URL адреси на 67% от общия брой предприятия, като има възможност за подобрение чрез адаптиране на по-добър списък с жълти страници и интернет каталози, които трябва да отпаднат предварително. Също така трябва да се има предвид, че съществуват разлики между очакваните полета за данни от италианския софтуер и полетата на файла от бизнес регистъра на НСИ - например кодът на областта и телефонният номер на предприятието са свързани в българския регистър за разлика от тяхното отделно използване в софтуера на ISTAT. Тази ситуация също допринася за сравнително ниския процент на съвпадение на резултатите вследствие на прилагането на българския и италианския софтуер.

След успешното завършване на практическата реализация, могат да се направят няколко **заключения:**

- По отношение на методологията се налага изводът, че добрият списък от бизнес каталози е основа за по-добри прогнози. Използването на точен брой полета с данни, необходими на софтуера, е основа за по-добро оценяване на предложените URL адреси.

- За ИТ технологиите може да се отчете, че използваната от ISTAT версия на Apache Solr е най-добрата, за да работи софтуерът подходящо, защото с последната актуална версия 6.5 на Apache Solr възникват технически проблеми.
- Констатирано е, че няма правни ограничения, защото са използвани публично достъпни ИТ инструменти и продукти (Bing Search APIs), за да се получат URL адресите на предприятията.

Заключение

Проведеното емпирично изследване „Извличане на информация от интернет за характеристики на предприятията (web-scraping)“ е първо в Big Data практиката на Националния статистически институт. Приложените методи и техники за набиране, структуриране, обработка и анализ на тази информация предопределя неговата уникалност. Като основен резултат от работата по четирите „сценария“ в контекста на SGA-I се открояват няколко ключови бъдещи предизвикателства, свързани с прилагането на техниката „web-scraping“ за извличане на характеристики на предприятията от техните уебсайтове и текущите дейности на екипа относно работата по SGA-II.

На **първо място,** отклоненията в крайните резултати са релевантни на стандартните отклонения в статистическата теория. Например при изпълнение на „use-case 1“ за автоматично намиране на URL адреса на предприятията от съвкупността стана ясно, че някои от фирмите са по-малко склонни да имат корпоративен уебсайт от други. Друг пример за отклонение е работата по „use-case 2“, където се оказа, че уебпаяка има способността да надценява идентифицирането на е-търговци и ако тези данни се използват директно за производство на официална статистика, може да има значително отклонение от действителността, без да могат данните да бъдат ажустирани. Освен това се установи, че при някои уебсайтове, които основно използват Java script, възниква техничеки проблем, при който е трудно да се извличат данни чрез „web-scraping“ и това също води до непълнота на данните. Основно предизвикателство в тези случаи ще бъде да се анализират отклоненията и да се разработи метод, чрез който събраните данни от интернет да се коригират по начин, който позволява тяхното използване за оценки, класифицирани като официална статистика. Това вероятно ще изисква методологическа работа около комбинирането на извлечени данни, данни от статистически изследвания или административни данни.

На **второ място,** предизвикателства съществуват и по отношение на етичните въпроси. Все повече физически лица и предприятия са достъпни в онлайн форма чрез интернет източници. Това е значителна възможност, но и предизвикателство, което е вероятно да доведе до по-голяма загриженост сред обществеността относно начина, по който правителството и държавните институции използват онлайн данните. В конкретното изследване НСИ извлича публични текстови данни от фирмените уебсайтове, за да идентифицира е-търговците или присъствието на дадено предприятие в социалните мрежи. В отговор на това НСИ и Евростат ще имат нужда да разработят прозрачни политики за прилагане на „web-scraping“, за да смекчат или изобщо да отпаднат обществените опасения относно събирането и използването на тези Big Data.

На **трето място,** използваните методи за „web-scraping“ в настоящото емпирично изследване са изцяло базирани на методи за събиране на текстова информация от уебстраници в интернет пространството. Това може да се превърне в голямо предизвикателство, тъй като интернет технологиите се развиват бързо, все повече и повече данни могат да бъдат кодирани в нестандартни форми, които са трудни за извличане - аудио- или видеофайлове, засилено използване на интерактивно или потребителско специфично съдържание. Ето защо, някои от уебсайтовете - по-специално на големите предприятия или тези, работещи в творческите индустрии, потенциално могат да станат трудно достъпни и невъзможни за извличане на информация. Поради тази причина е възможно да се наложи повторно извършване на „web-scraping“ на данни от висококвалифицирани специалисти във или извън НСИ, които ще изискват и по-голям финансов ресурс.

Практическата реализация на проекта поставя редица въпроси пред специалистите от НСИ, свързани с необходимостта от конструиране на сложните данни, извлечени в голям мащаб, преди по-нататъшен анализ, като прилагането на машинно обучение за валидиране, свързване и интегриране на данни. Не по-малко важен е проблемът с извършването на „web-scraping“ от онлайн машини в мрежата и прехвърлянето на извлечените данни от офлайн мрежи в онлайн хранилища. НСИ се нуждае от сходни системи за съхранение на данни, които могат да управляват целия жизнен цикъл на данните - съхранение, проектиране и поддържане на големи бази.

В заключение може да се каже, че разглеждайки Big Data с тяхната реална стойност и значимост и връзките им с други явления, смисълът на тяхното използване като един от възможните източници в официалната статистика изглежда по-скоро логичен, отколкото необичаен. Националните статистически институти трябва да имат фундаментални знания и да разширяват опита си по отношение на използването на Big Data в ежедневната статистическа практика и извън

нея. Прилагането на принципа „количество над качество“, възприет от потребителите на Big Data, не трябва да се пренебрегва. Дори когато източниците на Big Data не се използват за получаване на нови статистически продукти, те биха могли да се разглеждат като ефективно средство за намаляване на натовареността на респондентите, при условие, че методологичните предизвикателства могат да бъдат разрешени. Използването на Big Data за съставяне на ранни показатели за важни статистики като например данни за цените или бизнес цикъла е достатъчно сериозна опция. Прилагането на Big Data за краткосрочни прогнози също не е за пренебрегване.

В крайна сметка полетата на теорията и практиката са обединени от една цел - получаване на реални и навременни изводи от публично и лесно достъпни данни. Развитието на обществото и информационните технологии разкриват нови потребителски очаквания, обекти и феномени на интерес, за които официалната статистика не е в състояние да предостави данни. „Големите данни“ са съществена част от това развитие. В този смисъл може да се добави, че чрез използване на принципите на нанотехнологиите Big Data ще могат да се метрират и това ще даде облика на 21-ви век.

ИЗПОЛЗВАНА ЛИТЕРАТУРА:

Ангелова, П. (2013). Статистика. Свищов, СА „Д. А. Ценов“.

Богданов, Б., Г. Статева (2017). Бъдещето на изследванията и изследванията на бъдещето: възможни приложения на големите данни при производството на статистическа информация. Статистика, № 1.

Европа 2020 (2010). Стратегия за интелигентен, устойчив и приобщаващ растеж. Европейска комисия, Брюксел, 3 март, COM(2010) 2020 окончателен. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2010:2020:FIN:BG:PDF.

Alpaydın, E. (2004). Introduction to Machine Learning (Adaptive Computation and Machine Learning). MIT Press, ISBN 0-262-01211-1; Witten, I., Frank, E., Hall, M. (2011). Data Mining: Practical machine learning tools and techniques. 3rd ed. Morgan Kaufmann. ISBN-13: 978-0123748560.

Boeing, G., P. Waddell (2016). New Insights into Rental Housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listings. Journal of Planning Education and Research, 23 august. http://journals.sagepub.com/doi/pdf/10.1177/0739456X16664789; Vargiu, E., Urru, M. (2013). Exploiting web scraping in a collaborative filtering- based approach to web advertising. Artificial Intelligence Research, vol. 2, №1. http://www.sciedu.ca/journal/index.php/air/article/view/1390.

Buelens, B., P. Daas, J. Burger, M. Puts, J. Brakel (2014). Selectivity of Big data. Discussion Paper, №11, Statistics Netherlands. file:///D:/2014-11-x10-pub.pdf.

Daas, P., M. Puts, B. Buelens, P. van den Hurk (2015). Big Data as a Source for Official Statistics. Journal of Official Statistics, Vol. 31, No. 2, pp. 249-262. http://dx.doi.org/10.1515/JOS-2015-0016.

Eurostat (2015). Experiments for using big data in official statistics. 2015 UNECE Project.