Metody automatycznego pozyskiwania danych z Internetu dla celów biznesowych i cyberbezpieczeństwa

Raport

Zespół Aigocracy Institute

Poznań, 14.03.2020

Streszczenie kierownicze

Zawartość raportu i jego cele

Celem raportu jest przedstawienie przypadków użycia źródeł danych dostępnych w Internecie do wsparcia procesów biznesowych, marketingowych oraz związanych z cyberbezpieczeństwem w sektorze bankowym. Raport omawia podstawowe zagadnienia związane z technicznym pozyskiwaniem danych (np. opis wybranych narzędzi), jak również ujęcie ekonomiczne (np. możliwe do uzyskania efekty skali, szacunkowe koszty wykorzystania wybranych źródeł) i prawne (bieżące rozwiązania legislacyjne, licencje źródeł danych).

Wnioski

1. Narzędzia do automatycznego pobierania treści ze źródeł internetowych są wysoce personalizowane (ściśle dostosowane do wybranego źródła), decyzję o stworzeniu takiego narzędzia powinna więc poprzedzać wnikliwa analiza potrzeb informacyjnych oraz struktury, charakterystyk i jakości możliwych do ich zaspokojenia źródeł.
2. Przegląd źródeł danych wykazał, że wiele ze źródeł związanych z sektorem bankowym ma postać nieustrukturyzowaną lub półstrukturyzowaną, co utrudnia ich automatyczne przetwarzanie.
3. Źródła związane z cyberbezpieczeństwem zawierają najczęściej dane nieustrukturyzowane: są to głównie newsy na portalach branżowych lub komentarze na forach i w mediach społecznościowych. Widoczna jest potrzeba stworzenia bazy zagrożeń związanych z sektorem bankowym (podobnej do baz gromadzących luki bezpieczeństwa w oprogramowaniu, np. https://nvd.nist.gov/vuln/search).
4. Agregacja danych pochodzących z różnych źródeł może umożliwić podwyższenie jakości gromadzonych danych (np. dzięki weryfikacji tych danych w różnych źródłach lub uzupełnianiu brakujących elementów).
5. Kluczowym ograniczeniem automatycznego pobierania danych z Internetu są przepisy prawa i licencje obejmujące źródła.
6. Na rynku dostępnych jest wiele narzędzi służących do budowania web scraperów. Są to często narzędzia darmowe. Wybór konkretnego uwarunkowany jest (1) posiadanymi zasobami technicznymi i ludzkimi, (2) rodzajem i strukturą źródła, z którego mają być pobierane dane.
7. Ponieważ koszty opracowania i utrzymania narzędzi do automatycznego pobierania danych z wielu źródeł będą rosły szybciej niż liniowo przy dodawaniu kolejnych źródeł (ze względu na konieczność integracji danych z różnych źródeł będą rosły oraz ze względu na rosnące wymagania infrastrukturalne), może być zasadne stworzenie narzędzia służącego do pozyskiwania, agregacji i dostarczania danych różnym podmiotom z sektora bankowego. Pozwoliłoby to na uzyskanie efektów skali.

Słowa kluczowe

Ekstrakcja danych, pozyskiwanie danych, źródła danych, web scrapping, web crawling

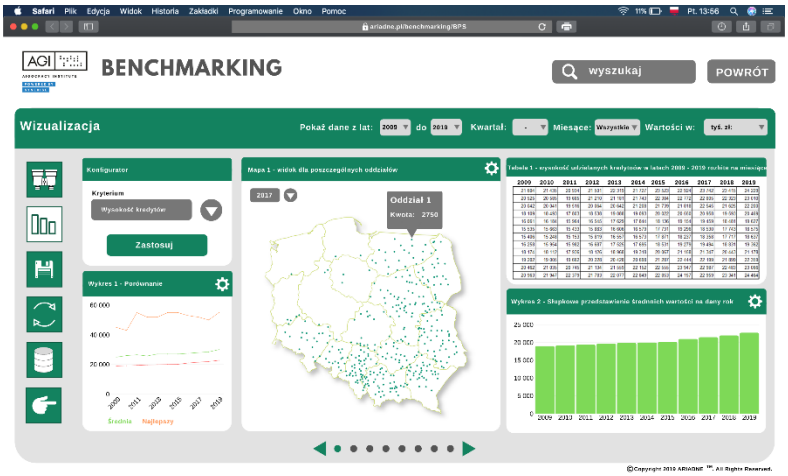
Rozdział 1
Wstęp

Zjawisko globalizacji w połączeniu z rozwojem Internetu przyczyniło się do powstania nowego rodzaju gospodarki, w której dane i informacje stanowią m.in. czynnik produkcji, zasób, produkt, a przedsiębiorstwa wykorzystują zdobyte informacje do budowy przewagi konkurencyjnej (Abramowicz, 2008)(Oleński, 2001). Rozwój technologii informacyjnych przyczynił się również do istotnych zmian modeli biznesowych opartych na zarządzaniu zbiorami danych oraz przetwarzaniu informacji, np. model infomediary (polegający na gromadzeniu oraz dostarczaniu klientom wyspecjalizowanych zbiorów danych) lub dostawcy treści (koncentrujący się na wytwarzaniu nowych danych lub informacji i dostarczaniu ich klientom). Wiele zbiorów danych jest dostępnych w postaci Open Data (tzw. otwarte dane, które charakteryzuje możliwość dowolnego, bezpłatnego wykorzystania). Inne udostępniane są na komercyjnych zasadach. Ponadto, możliwe jest pozyskiwanie danych publikowanych w formie tekstowej na stronach internetowych – taki proces nazywany jest z języka angielskiego web scraping (dosłownie: zdrapywanie danych).

Web scraping pozwala na zautomatyzowanie procesu pozyskiwania danych z witryn internetowych. Ręczne pobieranie danych polega na nawigowaniu po witrynie przez człowieka i kopiowaniu określonych fragmentów tekstów - taki proces jest czasochłonny i kosztowny a jego możliwości są ograniczone, bowiem człowiek, w przeciwieństwie do narzędzi web scrapingowych, nie jest w stanie przetwarzać wielu stron jednocześnie, zachowując przy tym odpowiednie tempo pracy.

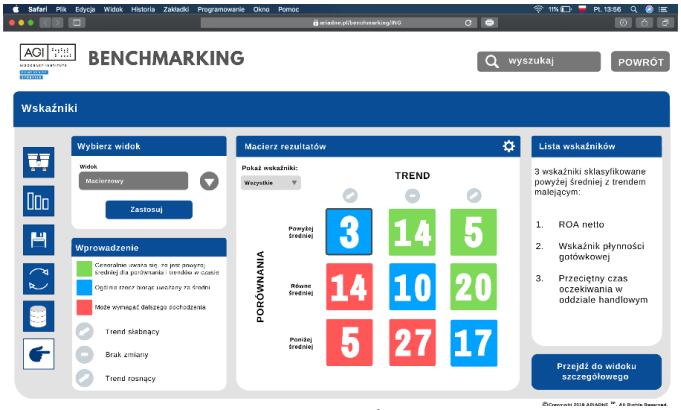
Kolejnym krokiem jest przekształcenie danych w informację, tj. nadanie im kontekstu, który pozwoli na sformułowanie wniosków. Informacje dają przedsiębiorstwom pełniejszy obraz rynku i potrzeb klientów, pozwalają podejmować trafniejsze decyzje, a w konsekwencji budować zasoby wiedzy. Chęć oszczędności czasu, automatyzacji ludzkiej pracy i zwiększenia dostępu do danych przyczyniły się do powstania pierwszych narzędzi do ekstrakcji danych internetowych. Web scraping, jako technika, służy do automatyzacji pozyskiwania treści z Internetu. Scraper znajduje określone elementy na stronie i zachowuje je w bazie danych, bądź w uprzednio zdefiniowanych formatach plików, jak np.: csv, czy xls. To, co dziś wymagałoby ogromu pracy człowieka, a przede wszystkim - zajęłoby dużo czasu, możliwe jest do wykonania automatycznie przez dedykowane narzędzia. Pomimo że web scraping nie jest nową technologią, scrapery nigdy wcześniej nie były tak powszechnie używane, jak dzisiaj. Pozyskane dzięki nim dane, przekształcone następnie w informacje, mogą stanowić o przewadze konkurencyjnej danej firmy. Przekształcenie danych w informację, tj. nadanie im kontekstu, pozwala na sformułowanie wniosków. Informacje dają przedsiębiorstwom pełniejszy obraz rynku i potrzeb klientów. Dane uznawane są za jeden z najcenniejszych zasobów, jakie może posiadać firma (Wiliams, 2018).

Mockup zaprezentowany na Rysunek 2**Błąd! Nie można odnaleźć źródła odwołania.** przedstawia porównanie wybranego banku pod względem



Rysunek 1 Przykładowy mockup. Źródło: opracowanie własne

wysokości udzielonych kredytów. Zestawienie banku ze średnimi wynikami sektora oraz z liderem dobrze obrazuje pozycję instytucji. Na mapie są przedstawione wszystkie oddziały zlokalizowane w



Rysunek 2 Przykładowy mockup. Źródło: opracowanie własne

Polsce. Tabela z prawej strony obrazuje wysokość udzielonych kredytów w poszczególnych miesiącach na przestrzeni kilku lat, a wykres poniżej przedstawia łączną wartość kredytów udzielonych przez bank w kolejnych latach.

Mockup zaprezentowany na Rysunek 3**Błąd! Nie można odnaleźć źródła odwołania.** pokazuje macierzowe porównanie kilku instytucji bankowych. Po lewej stronie znajduje się legenda do odczytu wyniku z macierzy rezultatów znajdującej się pośrodku. W wierszach od góry zaprezentowane są wyniki powyżej, równe i poniżej średniej wszystkich banków. W kolumnach natomiast jest zaprezentowany trend, od lewej malejący, stały i rosnący. Wynika z tego, że najkorzystniejsze wyniki zaznaczone są na zielono, a najmniej korzystne na czerwono. Okno znajdujące się po prawej stronie zawiera listę z wybranymi wskaźnikami, w przykładzie są one powyżej średniej z trendem malejącym.

3.3 Portale branżowe

Portale branżowe mają na celu publikację newsów ze świata finansów oraz informacji o sytuacji na rynku nieruchomości i trendach technologicznych, ze szczególnym uwzględnieniem ich wpływu na sektor finansowy. Wiele z takich branżowych portali, chcąc podążać za nowymi technologiami i rozwiązaniami, skupia się również na publikowaniu informacji związanych z cyberbezpieczeństwem, o zmianach w przepisach prawnych wraz z określeniem ich istotności dla przedsiębiorstw czy wybranych sektorów rynkowych, jak również powiadomienia o klęskach żywiołowych, katastrofach oraz epidemiach.

Portale branżowe publikują informacje o podobnej tematyce, można jednak wyznaczyć kilka znaczących, z punktu widzenia pobierania danych, aspektów, które je różnicują.

Po pierwsze, portale można klasyfikować pod względem struktury technicznej stron. Pomimo faktu, iż strony pozornie wydają się być zbliżone wyglądem, to ich struktura jest zupełnie inna. Pobieranie z nich danych wymagałoby więc dostosowania scraperów i opracowania reguł ekstrakcji pod konkretne źródło. Ma to swoje uzasadnienie w działaniu scraperów, które zostało opisane w dalszej części raportu (patrz **Błąd! Nie można odnaleźć źródła odwołania.**).

Po drugie, format pobieranych danych może okazać się trudny do dalszego przetwarzania. Pozyskiwanie informacji z opisywanych portali wymaga bowiem pobierania całych artykułów, czyli ciągu tekstu, które następnie muszą zostać poddane dalszej analizie, np. metodami przetwarzania języka naturalnego (NLP). Niekiedy portale kategoryzują publikacje, tak aby ułatwić ich wyszukiwanie, co może być wykorzystywane również przez roboty internetowe do identyfikowania treści określonego typu. Przykładem takiej strony jest ObserwatorFinansowy.pl, na której każdy artykuł został oznaczony odpowiednimi tagami, pomocnymi przy selekcji treści do pobrania.

Treści zamieszczone na portalach są obwarowane różnymi zapisami licencyjnymi lub innymi, które określają, w jaki sposób (i czy w ogóle) można je przetwarzać. Przykładowo, ObserwatorFinansowy działa na otwartej licencji (otwarta licencja zezwala na korzystanie z udostępnianych zbiorów pod warunkiem uznania autorstwa, czyli oznaczania źródła pochodzenia), jednak wiele innych portali nie udziela informacji o możliwości pobierania zasobów. W takich okolicznościach, przed rozpoczęciem scrapowania należy skontaktować się z odpowiednim działem danego portalu z prośbą o ustosunkowanie się do tej kwestii (patrz rozdział **Błąd! Nie można odnaleźć źródła odwołania.**).

Przykładowe portale branżowe znajdują się w Tabela 1.

Tabela 1 Przykładowe portale branżowe. Źródło: opracowanie własne.

Nazwa	Potrzeby informacyjne
Obserwator finansowy **(https://www.obserwatorfinansowy.pl/)**	• trendy technologiczne w sferze finansów oraz ich odbiór przez potencjalnych klientów; • monitorowanie informacji o naruszeniach bezpieczeństwa banków i ich klientów; • nowe regulacje i przepisy prawne; • trendy na rynku nieruchomości; • informacje o klęskach żywiołowych, katastrofach oraz epidemiach.
PwC **(https://www.pwc.pl/pl/branze/bankowosc-ubezpieczenia.html)**	• trendy technologiczne w sferze finansów oraz ich odbiór przez potencjalnych klientów; • trendy na rynku nieruchomości;
ale Bank **(https://alebank.pl/)**	• trendy technologiczne w sferze finansów oraz ich odbiór przez potencjalnych klientów; • nowe regulacje i przepisy prawne; • trendy na rynku nieruchomości;
CyberDefence24 **(https://cyberdefence24.pl/biznes-i-finanse/)**	• trendy technologiczne w sferze finansów oraz ich odbiór przez potencjalnych klientów; • monitorowanie informacji o naruszeniach bezpieczeństwa banków i ich klientów;
Forbes **(https://www.forbes.pl/rynek-nieruchomosci)**	• trendy w branży budowlanej