DOI: 10.15199/13.2016.6.29

Zastosowania technologii Big Data

Applications of the Big Data technology

mgr KAROL NOWAKOWSKI, Bank Gospodarstwa Krajowego

prof. dr inż. WOJCIECH NOWAKOWSKI, Instytut Maszyn Matematycznych, Warszawa

Streszczenie

Omówiono niektóre z zastosowań technologii Big Data, która jest jednym z najważniejszych wyzwań współczesnej informatyki. Wobec zmasowanego napływu wielkich ilości informacji obecnych czasach pochodzących z różnych źródeł, konieczne jest wprowadzanie nowych technik i technologii analizy.

Słowa kluczowe: Big Data

Abstract

Same of the Big Data applications is described. In view of the massive influx of large amounts of information nowadays from various sources, it is necessary to introduce new data analysis techniques and technologies.

Keywords: Big Data

Big Data w praktyce stosuje się już w wielu krajach, w wielu branżach i do wielu modeli biznesowych [1, 2, 3]. Dla przedstawienia najistotniejszych przykładów zastosowań zostały one zgrupowane według kryterium korzyści, jakie dane zastosowanie niesie. Poniżej zostaną pokrótce omówione dwie grupy przykładów, prezentujących *Big Data* jako narzędzie:

- Służące wzrostowi wydajności procesów biznesowych, całej organizacji czy też wydajności z punktu widzenia klienta,
- Redefiniujące istniejące branże lub tworzące nowe branże zmieniające rynek.

Big Data jako narzędzie służące wzrostowi wydajności

Najbardziej oczywiste i jednocześnie najczęstsze zastosowanie *Big Data* to poprawa wydajności. Dwa przykłady dotyczą modelu biznesu, w którym firma chce zapewnić wzrost wydajności swoim klientom, zaś dwa następne pokazują wzrost wydajności wewnętrznych procesów biznesowych. Ponadto zostaną przedstawione przykłady mechanizmów *Big Data* w przewidywaniu cen biletów lotniczych, przewidywaniu opóźnień na lotniskach spowodowanych pogodą, przewidywaniu potrzeb płatniczych osób płacących kartami oraz przewidywaniu potrzeb zakupowych klientów sieci hipermarketów.

Farecast

Farecast to założona w 2004 r. przez Orena Etzioni usługa, której celem było przewidywanie zmian cen biletów na połączenia lotnicze pomiędzy lotniskami w Stanach Zjednoczonych. Od kwietnia 2014 r. o usłudze tej należy mówić już w czasie przeszłym, ponieważ od momentu wykupienia Forecast w 2008 r. przez Microsoft usługa nie była rozwijana, aż do finalnego jej wyłączenia w roku 2014.

Farecast było pierwszą usługą, która szerokiemu gronu odbiorców nie zajmujących się informatyką czy problemami technologii informacyjnej uświadomiła na czym polega korzyść z *Big Data* [4].

Oprogramowanie, stworzone przez Orena Etzioni miało w regularnych odstępach czasu odwiedzać serwisy internetowe, na których znajdowały się ceny biletów lotniczych. W trakcie pierwotnego projektu badawczego program pobrał 12 tysięcy pozycji cenowych z wielu stron internetowych w ciągu kolejnych 41 dni. Aby uzyskać informacje o dynamice cen program powtarzał analizę codziennie, każdego dnia zdobywając lub weryfikując te same strony. Po kilkunastu miesiącach funkcjonowania usługa Farecast była w stanie określić jak zmieni się trend cen połączeń – ich długofalowe zmiany w funkcji czasu. Wraz z wydłużaniem czasu trwania analizy oraz jej głębokości (poprzez dodawanie nowych stron internetowych) stało się możliwe oddzielenie trendu od krótkoterminowych wahnięć, takich jak święta czy weekendy.

Podejście naukowe do analizy cen biletów lotniczych zmusza nas do odszukania korelacji pomiędzy czynnikami wpływającymi na cenę a jej zmiennością. Mantin B. oraz Gillen D. [5] wzięli pod uwagę w swojej pracy takie zmienne, jak m.in. liczba podróżnych na danej trasie (jako miernik dzisiejszej wielkości rynku), udziały linii lotniczych w ruchu na danej trasie obliczone w oparciu o HHI (wskaźnik pozycji konkurencyjnej), długość trasy (jako pochodną kosztu realizowania przelotu na danej trasie), populacje miast końcowych trasy (jako miernik potencjalnej wielkości rynku), dochód na mieszkańca w miastach końcowych (jako pochodną krańcowej skłonności do zakupu biletu lotniczego) i wiele innych.

Puller S.L. oraz Taylor L.M. [6] wskazali, że dobór czynników to tylko jeden z wymiarów analizy. Istotnym (o ile nie istotniejszym) jest zdaniem autorów analizowanie jedynie tych cen, po których została zrealizowana faktyczna transakcja, a nie wyłącznie cen oferowanych. Ceny oferowane nie podlegają według słów autorów takim samym prawom ekonomii jak ceny transakcyjne.

Autorzy algorytmu Hamlet [7], który był wykorzystywany w usłudze Farecast zaprezentowali podejście odmienne – narzędziowe. Nie próbowali poznać przyczyn zmienności cen, a jedynie sprawdzić czy wykorzystanie narzędzia *data mining*

(masowego zdobywania danych) z łatwo dostępnych źródeł internetowych pozwoli na obniżenie kosztów przelotów. Jedynym czynnikiem, jaki był w ich modelu analizowany, to upływ czasu. Model funkcjonujący przez 41 dni pobierał rzeczywiste dane z wielu stron internetowych, a aktorzy modelu – fikcyjni klienci – dokonywali zakupów biletów lotniczych. Zakupy te były rejestrowane dwutorowo – od razu po pojawieniu się w modelu oraz po oczekiwaniu, którego długość wyznaczały sugestie algorytmu.

Dla łącznie 4488 fikcyjnych klientów średnia oszczędność w cenie biletu wyniosła 4.4%. Dla podgrupy 341 klientów, którzy na sugestiach algorytmu skorzystali najbardziej, średnia oszczędność wyniosła 23.8%. Całkowita oszczędność w analizie wyniosła ponad 198 tys. dolarów, czyli średnio 44 dolary na jednego klienta. Po zakończeniu podstawowej analizy przeliczono również, ile oszczędziliby fikcyjni klienci, gdyby potrafili w pełni skutecznie przewidywać przyszłość kształtowania się cen. Wynik tak przeprowadzonych obliczeń dał wartość ponad 320 tys. dolarów całkowitej oszczędności, co oznacza, że skuteczność analizy opartej o algorytm Hamlet wyniosła 61.8%.

Miliony użytkowników serwisu Farecast nie musiało znać tych analiz by nabyć bilety oszczędniej niż przed uruchomieniem tego serwisu. Podobnie, jak twórcy algorytmu nie musieli rozumieć dlaczego ceny ulegają zmianom. Ponad sześćdziesięcioprocentowa skuteczność algorytmu w stosunku do jasnowidztwa osiągnięta została wyłącznie przez analizę jednego czynnika – czasu. Czynnika, który nie jest ani jedynym, ani nawet podstawowym czynnikiem, który służy firmom lotniczym do ustalania cen biletów.

Dane, które firma Farecast używała do swojej analizy, były bezpłatne i dostępne bez ograniczeń dla każdego internauty. Podczas gdy pozostałe przytoczone badania bazowały na danych, do których nie ma w zwyczajnych warunkach dostępu, algorytm Hamlet korzystał z całej populacji danych i to danych ogólnodostępnych.

Konstatacja ta jest o tyle istotna, że uświadamia problem użytkowy. Nawet gdyby w ramach pozostałych dwóch badań skuteczność przewidywania algorytmu byłaby wyższa (nie jest ona w opisach tych badaniach podawana) to ceną za to osiągnięcie byłoby pozbawienie algorytmu możliwości rzeczywistego wykorzystania biznesowego.

Poza środowiskiem badawczym nie byłoby przecież możliwości bieżącego zasilania algorytmu rzeczywistymi danymi o zapotrzebowaniu, klientach czy o kwotach poszczególnych transakcji.

Sukces Farecast polegał nie tyle na skuteczności, co na prostocie koncepcji – zrozumiałej dla wszystkich użytkowników zasadzie, dzięki której potencjalni klienci nie zakładali, że są oszukiwani przez kolejną korporację, która pod przykrywką badań naukowych chce im coś sprzedać na gorszych niż by sobie życzyli warunkach.

Od roku usługa Farecast już nie funkcjonuje. Po wykupieniu jej przez Microsoft stała się częścią wyszukiwarki Bing jako „Price Predictor". Aktualnie Microsoft rozwija projekt Kayak, Google zaś, w którym pracuje Orena Etzioni, inwestuje w narzędzie Google Flights.

FlyOnTime.us

Dla tej samej branży ważnym narzędziem jest też projekt *FlyOnTime.us.*

Serwis ten stara się udzielać odpowiedzi na inne związane z podróżami lotniczymi pytanie: „Ile będę zmuszona/y czekać na lotnisku".

Podobnie jak w przypadku poprzedniej usługi również i tutaj autorzy algorytmu nie postawili na zrozumienie związków przyczynowo-skutkowych pomiędzy zmianą pogody, a opóźnieniami. Ich celem było przewidzenie możliwych opóźnień poprzez szukanie korelacji danych, nie zaś tworzenie wzoru matematycznego, którego argumentami będzie mgła, wielkość lotniska czy średnioroczna temperatura przy ziemi.

Po wprowadzeniu danych i uruchomieniu algorytmów szukających korelacji okazało się, że najbardziej niebezpiecznym terminem podróży dla dziesięciu lotnisk w Stanach Zjednoczonych, które wg Air Carrier Activity Information System (ACAIS) obsłużyły najwięcej pasażerów w 2013 r. jest 26 grudnia. W tym dniu odwoływanych jest najwięcej lotów. Nawet tornada powodują odwołanie mniej lotów z tych lub do tych lotnisk.

Jeżeli jednak lot zaczyna się na lotnisku w Atlancie

Liczba odwołanych lotów w wybranych amerykańskich portach lotniczych. The number of canceled flights at selected US airports

	ATL	LAX	ORD	DFW	DEN	JFK	SFO	CLT	LAS	PHX
Zwykły dzień	2	1	3	2	1	3	2	1	1	1
Dzień po Memorial Day	1	1	1	0	1	0	1	1	0	0
Memorial Day	1	0	**11**	1	1	2	1	3	0	1
Dzień po Memorial Day	1	1	2	2	0	12	2	1	1	1
Dzień pracujący	1	0	1	0	1	0	0	1	0	0
Środa przed Św. Dziękczynienia	1	0	1	0	1	1	0	0	0	0
Święto Dziękczynienia	0	0	1	0	0	0	0	0	0	0
Piątek po Św. Dziękczynienia	1	0	0	0	0	1	0	0	0	0
Sobota po Św. Dziękczynienia	0	1	0	0	0	0	1	0	1	0
Niedziela po Św. Dziękczynienia	0	1	0	0	0	0	1	0	0	1
24 grudnia	2	0	1	1	1	0	1	0	0	1
25 grudnia	**56**	1	2	2	2	2	3	4	2	2
26 grudnia	**23**	**8**	**11**	**7**	**5**	**60**	**11**	**18**	**7**	**5**

ATL Hartsfield-Jackson, Atlanta, LAX Los Angeles, ORD O'Hare, Chicago, DFW Dallas–Fort Worth, DEN Denver, JFK John F. Kennedy, Nowy Jork, SFO San Francisco, CLT Charlotte, LAS McCarran, Las Vegas, PHX Sky Harbor, Phoenix.

(Hartsfield-Jackson, ATL), to wyjątkowo najmniej pewnym dniem jest 25 grudnia, z którego odwoływanych jest średnio 56% lotów. Szczegóły obliczeń przedstawiono w tabeli na stronie następnej. Analiza ta jest tylko jednym z wielu przekrojów, w omawianej usłudze. Czym jednak taka analiza różni się od tradycyjnych badań statystycznych?

Powyższa analiza jest tylko jednym z wielu przekrojów, w omawianej usłudze. Czym jednak taka analiza różni się od tradycyjnych badań statystycznych? Po pierwsze, do przygotowania analizy nie posłużyła reprezentacyjna próba lotnisk i połączeń tylko kompletne zestawienie wszystkich lotnisk i wszystkich połączeń. Dane pochodziły z dwóch zasadniczych źródeł: informacji o przelotach z Bureau of Transportation Statistics (agenda rządu USA ds informacji o transporcie) oraz National Oceanic and Atmospheric Administration (agenda rządu USA do spraw nadzoru nad oceanami i atmosferą), skąd pochodziły dane o pogodzie. Połączenie tych dwóch zestawów danych było niezbędne do udzielenia odpowiedzi na pytanie stawiane przez autorów FlyOnTime.us: jak pogoda wpływa na terminowość lotów. Dzięki ten analizie okazało się, że śnieg spowoduje średnie opóźnienie 24 minut jeżeli wylot następuje z lotniska w Phoenix (PHX), ale nie spowoduje średnio żadnego opóźnienia jeśli wylot następuje z lotniska Chicago O'Hare (ORD). Osiemdziesiąt pięć percent startów we mgle ma na lotnisku Denver International (DEN) opóźnienie 35 minut, ale ta sama statystyka dla lotniska Atlanta Hartsfield-Jackson (ATL) oznacza oczekiwanie o prawie godzinę więcej.

Dzięki analizie wszystkich połączeń lotniczych analiza subpopulacji statystycznej nie jest obarczona większym błędem niż przy wyborze nawet w pełni poprawnej próby losowej. Załóżmy, że próbą losową będzie 1030 lotów z prawie 10 milionów realizowanych w USA (czyli 0,0103%). Zgodnie z wcześniejszymi informacjami 56% lotów w dniu 25 grudnia z lotniska ATL jest odwoływane. W roku 2009 były to 894 loty w populacji 10 milionów. W próbie statystycznej znalazłoby się więc 0,092 lotu z tego lotniska w tym dniu. Nawet jeśli w próbie losowej znalazłoby się jeden czy dwa loty w tej konfiguracji, zostałyby one uznane za błąd statystyczny. To właśnie można zauważyć w analizie *Big Data*, a nie da się zauważyć w tradycyjnej analizie statystycznej.

Drugą różnicą jest zakres stawianych hipotez. W badaniu statystycznym stawiane są hipotezy, które następnie są weryfikowane zgodnie z zasadami statystyki matematycznej. W analizie *Big Data* hipotezy – nawet jeśli są – nie mają wpływu na prowadzone obliczenia. Celem autorów narzędzia było wykrycie wpływu pogody na terminowość lotów na poszczególnych lotniskach – stąd pierwotnie wgrany zasób danych do analizy (dane o lotach oraz o pogodzie). Automatyczna obróbka dokonana przez komputer wskazała, że dużo silniejsza korelacja niż z pogodą występuje wobec określonych dat. Obróbka danych źródłowych w *Big Data* nie polega więc nie na rozumieniu przyczyny zależności, ale jedynie na jej dostrzeżeniu.

Ze znalezieniem korelacji komputer poradzi sobie o znacznie lepiej niż człowiek. Komputer po prostu przeliczy miliony kombinacji danych szukając korelacji pomiędzy dowolnymi dwiema zmiennymi, dowolnymi trzema zmiennymi i tak dalej. W końcu komputer w wyniku swoich obliczeń poinformuje, że korelacja występuje tu i tu. I to na tej wstępnie wyselekcjonowanej liście czynników korelujących badacz może szukać związków przyczynowo-skutkowych.

Czy jednak jest to konieczne? I tak i nie. Potencjalny pasażer postara się tak zorganizować swoje święta, by omijać lotniska o największych opóźnieniach i najwyższym ryzyku odwołania lotu. Dla pasażera nie jest przecież istotna przyczyna tych negatywnych zdarzeń. Dla podmiotu zarządzającego lotniskiem poznanie przyczyny jest niezbędne. Dlaczego konkurujące z nami lotnisko radzi sobie z opóźnieniami lepiej niż my? Mając analizy Big Data zmienia się jednak treść tak postawionego pytania: *dlaczego lotnisko w Los Angeles ma we mgle maksymalne spóźnienia tylko 19 minut, a my w Atlancie mamy aż 88 minut?* Z menadżerskiego punktu widzenia takie pytanie ma już inny poziom trudności i mniejszą ilość potencjalnych przyczyn do zinterpretowania. Przy tym w czasie poświęconym na analizę konkurencji można skupić się na mniejszej liczbie aspektów i lepiej je opracować.

VISA

Właściciel marki kart płatniczych VISA funkcjonuje od 1970 r. W grudniu 2013 r. firma łączyła 14.5 tys. instytucji finansowych, a jej klienci korzystali z 2.2 miliarda kart płatniczych. W 2012 r., w samych tylko Stanach Zjednoczonych obsłużono 26.2 miliardów transakcji płatniczych oraz gotówkowych z użyciem kart VISA (szczegóły poniżej). Liczba ta jest większa o jedną trzecią jeśli uwzględnione zostaną transakcje z pozostałych części świata.

Obroty VISA Inc., właściciela marki kart płatniczych VISA
Turnover of VISA Inc., owner of the brand cards VISA

	2009	2010	2011	2012	2013
Kwota transakcji [mld $]	4423	5191	6029	6409	6970
Liczba transakcji [mld szt.]	62,2	70,8	77,6	81,6	89,7
Wydanych kart [mln szt.]	1808	1897	2011	2128	2219

Samo przeprowadzenie tych transakcji i uzyskanie od nich prowizji nie kończy możliwości osiągania dochodu firmy VISA. Historie transakcji są źródłem informacji o przyzwyczajeniach zakupowych klientów, o ich budżetach, o miejscach pobytu czy nawet o składzie rodziny, czy sytuacji materialnej. Takie dane pozwalają na lepsze dopasowanie kampanii reklamowej, lepszy dobór partnerów VISA do organizacji zakupów „kuponowych" czy po prostu zniżek. Decyzje te mogą odbywać się nie tylko na poziomie produktów, ale i na poziomie miejsc, w których odbywa się transakcja.

Takie dane trzeba jednak zanalizować. Trzeba dokonać ich komputerowej obróbki – trzeba tak te 36 TB transakcji filtrować, sortować i grupować, by uzyskać potrzebne odpowiedzi. Analiza danych prowadzona tradycyjnymi metodami statystycznymi wyrażonymi w dwóch krokach „postaw hipotezę; zweryfikuj hipotezę" powtarzana odpowiednio długo pozwalała określić segmenty klientów i ich preferencje zakupowe. Analiza zbiorcza danych z dwóch lat zajmowała miesiąc. Natomiast wprowadzenie narzędzia Hadoop firmy Apache wraz z rozwinięciem koncepcji w Map Reduce Google'a pozwoliło na skrócenie czasu analizy do 13 minut. A różnica w czasochłonności analizy przede wszystkim podnosi wydajność dotychczasowych procesów segmentacji klientów. Ale przy tak dużej skali poprawy należy mówić o podniesieniu wydajności całego pio-

nu marketingu, ponieważ możliwe jest częstsze przeszukiwanie bazy danych i częstsze poszukiwanie korelacji.

Konkurencyjna firma – Master Card, a dokładnie jej dział badawczy MasterCard Advisor – przeprowadził analogiczną analizę na swoich zbiorach danych. Program wskazał faktyczną korelację tam gdzie nikt jej nie szukał. Okazało się, że ludzie, którzy tankują benzynę około czwartej po południu w sąsiedztwie swojego miejsca zamieszkania, miejsca pracy lub drogi łączącej te miejsca, z bardzo wysokim prawdopodobieństwem w ciągu kolejnej godziny wydadzą między 35 a 50 dolarów w restauracji lub sklepie spożywczym.

Wiedząc, że taka korelacja istnieje trudno nie stwierdzić chociażby: „*przecież to oczywiste; dostaję pensję, uzupełniam bak i żołądek*”. Skoro jest to jednak takie oczywiste, to dlaczego nikt nie wskazał tej korelacji wcześniej? Tego typu gotowe wskazówki są warte bardzo dużo dla działów marketingu czy działów sprzedaży stacji benzynowych, restauracji i sklepów spożywczych, ponieważ podnoszą ich wydajność. Te firmy nie muszą już szukać szans – muszą jedynie leżące przed nimi szanse wykorzystać. Kupony promocyjne otrzymane na stacji do nieodległego sklepu spożywczego; odbiór zakupów spożywczych na stacji benzynowej; wejściówka i rezerwacja stolika od razu na stacji; podrzucenie na telefon komórkowy nowości, ofert dnia czy nowego menu restauracji wtedy kiedy czekamy na swoją kolejkę do kasy na stacji benzynowej – to tylko niektóre pomysły jakie można wdrożyć posiadając taką wiedzę.

Wycieczka w przyszłość

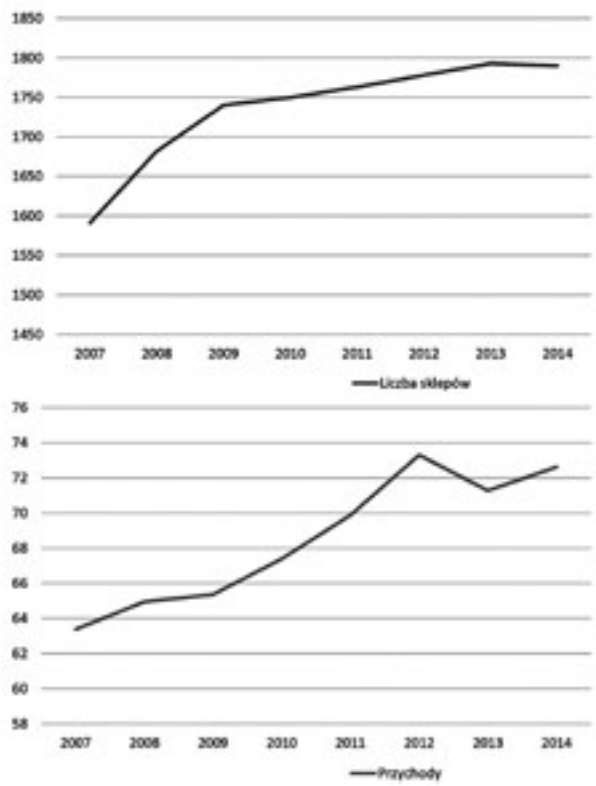
Dzisiejsze dane pozwalają nam zatem spojrzeć w przyszłość. Zakup wakacji to sugestia w jakim kraju będziemy korzystać z karty płatniczej. Zakupy w Świecie Dziecka czy Smyku to sugestia planów zakupowych na najbliższe kilkanaście lat. To działania poprawiające wydajność procesów zarządzania sprzedażą w firmach, a osiągnięcie korzyści nie jest zastrzeżone wyłącznie dla tych firm, które gromadzą wielkie ilości danych.

Sieć sklepów Target (*rys. 1*) umiejętnie gromadzi informacje o preferencjach zakupowych swoich klientów. Firma wykorzystuje do tego celu dwa źródła danych – listy prezentowe oraz szczegółowe historie transakcji osób, które korzystały z karty lojalnościowej, kuponów upominkowych lub kuponów rabatowych.

W drodze analizy odkryto 25 produktów, których odpowiednia korelacja pozwoliła na określenie z wysokim prawdopodobieństwem, że klientka jest w ciąży. Posiadając te dane możliwe było przeanalizowanie historii transakcji wstecz i określenie zmian w preferencjach zakupowych kobiet we wczesnych etapach ciąży. Jaki był efekt i skuteczność tych poszukiwań?

Charles Duhigg opisuje [8] historię jaka zdarzyła się w jednym ze sklepów Target w stanie Minnesota, do którego w stanie wielkiego zdenerwowania wpadł ojciec nastoletniej córki trzymając w ręku kopertę wysłaną na jej nazwisko zawierającą specjalnie wyselekcjonowane dla niej oferty ubranek dziecięcych i łóżeczek. Kierownik sklepu osobiście zobowiązał się do wyjaśnienia sprawy i po kilku dniach oddzwonił do klienta z wyjaśnieniem. Tym razem ojciec klientki powiedział „*Okazało się, że w moim domu doszło do pewnych zdarzeń, których nie byłem świadomy. Córka rodzi w sierpniu. Jestem panu winny przeprosiny*”.

Algorytmy szukające korelacji w transakcjach sklepowych przewidziały ciążę. Legalność takiego działania nie budzi wątpliwości, ponieważ klientka zgodziła się na przetwarzanie danych



Rys. 1. Liczba sklepów i przychody sieci Target. Wykres górny – liczba sklepów sieci Target w USA w latach 2007–2014. Wykres dolny – dochody sieci w tym samym okresie w bilionach dolarów. Na podstawie:
Fig. 1. Number of stores and revenues networks Target. Upper – the number of chain stores Target in the United States in the years 2007–2014. Lower - income networks in the same period in the trillions of dollars. Based:
http://www.statista.com/statistics/255957/revenue-of-target-in-north-america oraz http://www.statista.com/statistics/255965/total-number-of-target-stores-in-north-america/

o sobie i swoich zakupach przystępując do programu lojalnościowego sieci handlowej. Pojawia się tutaj aspekt etyczny – czy mimo, że działania sieci sklepów były zgodne z prawem, są one nieetyczne? Lub szczerzej – czy zbieranie, przetwarzanie i wyciąganie wniosków z danych, które dobrowolnie zostawiliśmy i nawet zgodziliśmy się co do ogólnego celu tej analizy (marketingu i promocji) jest etyczne bez względu na wnioski, które wyciągniemy i bez względu na to jak te wnioski wykorzystamy?

Jeżeli lista naszych zakupów wykazała, że naszą preferencją są bezglutenowe słodycze w piątki, a żelki w poniedziałki, i sklep zaproponowałby nam zniżkę na te produkty – to nie byłoby problemu. Ale jeżeli okaże się że jesteśmy chorzy na raka i zanim lekarze postawią diagnozę wyjmiemy ze skrzynki reklamy wszystkich okolicznych domów pogrzebowych? A jeżeli okaże się, że kwalifikujemy się z racji na korelacje na przyszłego przestępcę, to czy naszych sąsiadów i znajomych ucieszą zniżki na broń palną i kupon na szkolenie na strzelnicy, który zostanie nam zaproponowany z hasłem „*skorzystaj już dziś!*”? Sieć sklepów Target rozwiązała ten problem inaczej.

Gazetki ofertowych są personalizowane częściowo. Część jest standardowa dla każdego lub wielu odbiorców. Osoba w ciąży dostanie więc np. więcej ofert „dziecięcych” niż jej są-

siadka, ale obie gazetki będą podobne. Indywidualny przekaz reklamowy ukryty, a skuteczność przekazu zmniejszona po to by nie budzić obaw etycznych.

Im bardziej szczegółową przyszłość jesteśmy przewidzieć na podstawie wniosków z modeli korelacyjnych tym szybciej dochodzimy do powstania nowych informacji, nowych usług, a w końcu do nowych branż.

Big Data redefiniuje branże lub tworzy nowe dane

W tym ostatnim przypadku można stwierdzić, że ilość przechodzi w jakość. Namalowany pędzlem obraz konia pozwala nam stwierdzić jak ten koń wygląda. Nie dowiemy się jednak o różnorodności gatunków konia, bo zbyt dużo malarzy byłoby potrzebnych by taką wiedzę przekazać. Żeby to osiągnąć potrzebujemy wydajniejszego, szybszego tworzenia obrazu – potrzebujemy fotografii. Nawet jeśliby wykonanie zdjęć trwało jak niegdyś kilkadziesiąt minut, to i tak byłby to skok jakościowy w stosunku do malowania. Kiedy tworzenie obrazu zajmuje ułamek sekundy, możemy robić serię zdjęć ukazujących konia w biegu, Zmiana tylko ilości zdjęć powoduje, że uzyskujemy nową jakość informacji. Kamery szybko-klatkowe wykonują kilkaset ujęć na sekundę co pozwala dostrzec niewidzialne gołym okiem ruchy mięśni czy włosia. Pomijając to, że nowe zdobycze rejestracji obrazów – fotografii w kolorze czy trójwymiarowej – samo zwiększenie ilości zdjęć tworzy nową wartość samą w sobie. To właśnie jest cechą technologii *Big Data*, że sam fakt wzrostu ilości informacji zmienia całe branże. Oto przykłady:

Redefinicja branży tłumaczeń maszynowych

Tłumaczenia dokonywane przez maszynę są szybsze i tańsze niż te dokonywane przez ludzi. Co jednak istotniejsze, tłumaczeniami maszynowymi można przetłumaczyć setki dokumentów bez ograniczeń w liczbie pracowników, a jedynie z ograniczeniem mocy obliczeniowej komputerów, którą właściwie można dowolnie skalować.

Stąd też było to jedno z tych wyzwań, które spędzało sen z powiek twórców pierwszych komputerów lampowych już w latach 40'tych XIX wieku [9]. Pierwszym poważnym podejściem do tego zagadnienia był projekt IBM zapoczątkowany w latach 50-tych tego samego wieku. Po wprowadzeniu do bazy danych zawrotnej jak na ówczesne czasy liczby dwustu pięćdziesięciu par słów w języku angielskim oraz rosyjskim i jedynie sześciu reguł gramatycznych oprogramowanie IBM'a zdołało względnie poprawnie przetłumaczyć 60 zdań tekstu. Kierownik tego projektu, Leon Dostert, w przekonaniu o wielkim sukcesie wieszczył, że już za trzy lata tłumaczenie maszynowe będzie codziennością branży tłumaczeń. Jak wiemy – tak się nie stało. Program badawczy został zamknięty w 1966 r. Tłumaczenia były niezadowalające, a moc komputerów znacznie ograniczona. Nie było już miejsca na nowe słowa, reguły potrzebowały dopisywania nowych wyjątków – a przecież język jest tworem żywym i zbyt małe tempo aktualizacji słowników oznacza, że tłumacz maszynowy cofa się w swojej wiedzy.

Redefinicja kontroli poprawności pisowni

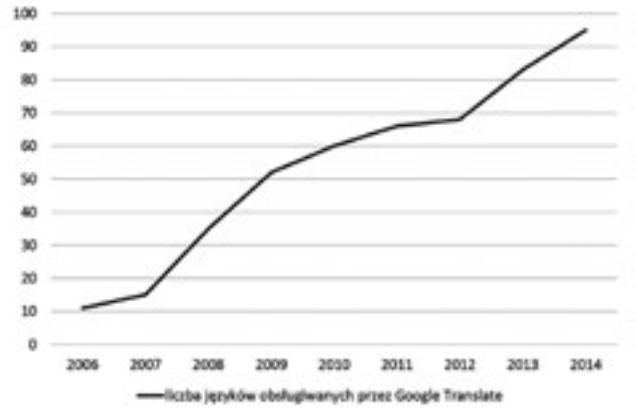
W latach 80. XIX w. IBM powrócił do projektu kontroli poprawności pisowni. Komputery od lat 50. zmieniły się prawie całkowicie (jedynie technologia dysków twardych pozostała właściwie niezmienna aż do dnia dzisiejszego), a ich moc obliczeniowa wzrosła o wiele rzędów wielkości. Pomysłem było wykorzystanie statystyki do określenia co oznacza słowo w danym kontekście, jakiego słowa program powinien użyć i jaki jest sens przekazywanego zdania. Nie skupiano się zatem na regułach gramatycznych i wyjątkach od nich, ale na statystyce wystąpień.

Takie oprogramowanie potrzebowało jednak tekstu dwujęzycznego przetłumaczonego w pełni poprawnie tak, by algorytm nie wprowadzał w błąd. Wykorzystywanym źródłem były transkrypcje parlamentu kanadyjskiego z okresu dziesięciu lat. Transkrypcje te były przygotowywane z najwyższą dbałością o szczegóły, a ponadto były bezpłatne i obszerne. Początkowo program odnosił sukcesy, jednak później nie notowano dalszej poprawy jakości tłumaczenia ze wzrostem zestawu tekstów.

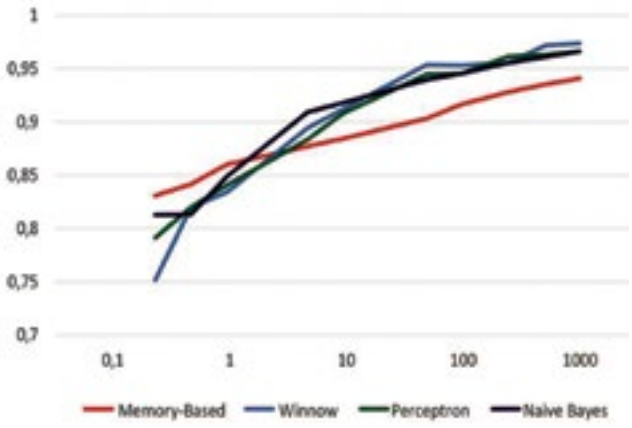
Obszerne i wysokiej jakości teksty były zbyt sformalizowane, język zbyt administracyjno-prawniczy, a ilość tekstów wciąż za mała. Choć zbiór liczący trzy mln zdań był kilka tysięcy raz większy od źródeł początkowych skala nie była jednak wystarczająca. Projekt został w 1996 r. zakończony z powodu braku perspektyw na dalszą poprawę jakości tłumaczeń.

Do koncepcji tłumacza powrócono w 2006 r. w firmie Google. Z racji rozwoju zdolności komputerów do magazynowania i przetwarzania informacji wprowadzono zmiany. Google Translate miał pobrać każdy dwujęzyczny tekst, na jaki natrafił w internecie. Było to naturalne zadanie dla giganta z Redmond – jest on do tej pory zarządcą największej bazy o stronach internetowych świata, w której przechowuje nie tylko informacje o nich, ale nawet i same strony. Google posiada na swoich dyskach miliardy zdań – wiele z nich jest dostępne dla wyszukiwarki również w formie przetłumaczonej. Mowa tu nie tylko o witrynach ONZ czy UE, ale również o olbrzymiej liczbie firmowych stron internetowych prowadzonych w kilku językach obcych.

Czy takie tłumaczenia są zawsze w pełni poprawne? Oczywiście, że nie. Ale ich olbrzymia ilość zastępuje jakość. Błędne tłumaczenia stanowią pewien procent wszystkich tłumaczeń – jeżeli jednak algorytm zauważa, że na dziesięć



Rys. 2. Języki obsługiwane przez Google Translate
Fig. 2. The languages supported by Google Translate
W pierwszym etapie został opracowany tłumacz z języka angielskiego na niemiecki. Do końca 2006 r. program obsługiwał 11 języków: angielski, arabski, chiński (uproszczony), francuski, hiszpański, japoński, koreański, niemiecki, portugalski, rosyjski i włoski, a obecnie już blisko 100. Tłumacz został udoskonalony o transliterację, syntezę mowy i słowniki frazeologiczne, o możliwość doboru słów i ocenę tłumaczenia.



Rys. 3. Zmiana wydajność mechanizmów korekty tekstu przy przyroście bazy słów
Fig. 3. Change of performance adjustment mechanisms text increment database of words Oś pionowa: dokładność testu, oś pozioma: miliony słów. Na wykresie: kolor czerwony – Memory-Based, kolor niebieski – Winnow, kolor zielony – Perceptron, kolor granatowy – Naive Bayes. Źródło: Banko M., Brill E., 2001.

tłumaczeń dziewięć tłumaczy tak, a jedno inaczej, to uznaje, że najprawdopodobniej te pojedyncze źródło jest błędne. Zauważmy przy tym, że korzystanie ze źródła jakim jest sieć internet nie pozwala również na starzenie się języka.

Dzięki obszerności źródeł internetowych powstają algorytmy tłumaczeń z wielu języków, co w efekcie tworzy sieć umożliwiającą tłumaczenie pomiędzy językami, w których nie istnieją bezpośrednie przekłady. Np. przetłumaczenie tekstu z języka urdu na język galicyjski nie stanowi już żadnego problemu

Redefinicja jaka nastąpiła w tym przypadku polega na rozszerzeniu branży tłumaczeń o nowe sektory. Zmiany te następowały już w XXI wieku, a ich pojawienie się stało się możliwe wyłącznie dzięki zdolności gromadzenia i analizowania wielkich zbiorów danych oraz właściwie nieograniczonej dostępności źródeł cyfrowych. Powstał sektor oprogramowania tłumaczeniowego CAT (*Computer Assisted/Aided Translation,* lub używając innej nazwy – MAHT czyli *Machine Assisted Humant Translation*). Oprogramowanie to nie zastępuje człowieka w tłumaczeniu tekstu, ale wspiera go w znaczący sposób. Wsparcie polega na tym, że komputer tłumaczy to, o czym jest przekonany, na określonym poziomie pewności. Resztę działania musi wykonać człowiek – musi nadać tłumaczeniu odpowiedni do przeznaczenia materiału nastrój czy charakter. Narzędzie Google Translate należy do jednego z takich narzędzi (choć oczywiście jest to przykład koncepcji CAT w zastosowaniu prywatno-amatorskim). Korzystając z usługi tłumacza Google nie ufamy mu w 100%, ale jego podpowiedzi znacząco ułatwiają nam zrozumienie tekstu napisanego w nieznanym nam języku obcym.

Oprogramowanie typu CAT nie istniało przed epoką masowego przetwarzania danych osobowych. Zwiększenie zdolności tłumaczeniowej komputerów stworzyło nowy sektor redefiniując branżę. Oprogramowanie typu CAT nie odebrało rynku zawodowym tłumaczom. Stało się inaczej – z jednej strony pozwoliło amatorom na zapoznawanie się z materiałami w obcym języku, co spopularyzowało potrzebę uzyskiwania tłumaczeń zawodowych, a CAT stało się narzędziem codziennej pracy tłumaczy. Ich praca jest wydajniejsza i prostsza. Podobna przemiana nastąpiła w niezbyt odległej pracy – korekcie tekstów.

Firma Microsoft, właściciel najbardziej popularnego na świecie edytora tekstów Office Word stanęła w 2000 r. przed zadaniem polegającym na poprawieniu jakości algorytmu sprawdzającego poprawność tekstów. Wówczas Word był absolutnym liderem wśród narzędzi służących do pisania tekstu, zwłaszcza w zakresie wykorzystania typowo biurowego. Jego algorytm „podkreślania na czerwono" był najlepszy z dostępnych na rynku – choć nie najlepszy w ogóle, ponieważ jego skuteczność sięgała od 75% do 84%. Poprawność w rejonie dolnej granicy tego przedziału oznaczała, że na cztery zanalizowane słowa jedno będzie niepotrzebnie podkreślone jako niepoprawne lub zostanie oznaczone jako zgodne z zasadami mimo, że będzie stanowiło błąd.

Początek XXI wieku to początek ery *Big Data*. Pracownicy firmy Microsoft postanowili sprawdzić, czy koncepcja ta może okazać się przydatna w rozwiązaniu ich problemu. Postanowili sprawdzić skuteczność trzech najczęściej wykorzystywanych algorytmów. Najprostszy miał skuteczność 75%, najbardziej skomplikowany osiągał wartość 84%.

W celu przeprowadzenia dalszych prób do wszystkich trzech algorytmów podłączono bazę kolejno dziesięciu milionów, stu milionów oraz ostatecznie miliarda słów. Efekt badania zaskoczył nie tylko badaczy, a także pozwolił przepowiedzieć rychłe zmiany w branży.

Każdy z algorytmów uzyskiwał po podłączeniu większej bazy słów wyższą poprawność (*rys. 3).* Najprostszy algorytm poprawił się o 20 punktów procentowych sięgając poziomu 95%. Algorytmy, które pierwotnie były lepsze, poprawiały się w mniejszym stopniu. Najbardziej złożony algorytm z pierwotnym wynikiem poprawności 84% osiągnął po podłączeniu miliarda słów tylko 94% poprawności. Dla pracowników Microsoft stało się jasne, że to nie algorytm stanowi jądro ich narzędzia, a olbrzymia baza danych słów.

Microsoft nie przewidział jednak jak poważny z punktu widzenia rynkowego jest to wniosek. Algorytm narzędzia można zaszyfrować, opatentować czy chronić na wiele innych sposobów. Bazy słów w danym języku nie można zastrzec, a koszt jej zdobycia to jedynie krótki czas potrzebny na skanowanie słownika. Programistom wyrwano kawałek rynku. Microsoft stracił przewagę konkurencyjną.

Dzisiaj sprawdzanie poprawności pisowni wbudowane jest w niemal każdą z popularnych przeglądarek internetowych dostępnych za darmo. Funkcja, która niegdyś dawała zarobek stała się w konsekwencji rozwoju *Big Data* funkcją powszechnie dostępną. To typowe zjawisko dla Internetu.

Literatura

[1] Nowakowski K., Nowakowski W.: 2016. Big Data *ante portas*. Elektronika – konstrukcje, technologie, zastosowania, nr 2/2016, str. 27–30, DOI: 10.15199/13.2016.2.5
[2] Nowakowski K., Nowakowski W.: 2016. Hadoop, narzędzie technologii Big Data i jego aplikacje. Elektronika – konstrukcje, technologie, zastosowania, nr 3/2016, str. 33–36, DOI: 10.15199/13.2016.3.7
[3] Nowakowski K., Nowakowski W.: 2016. Big Data czyli analiza korelacji. Elektronika – konstrukcje, technologie, zastosowania, nr 5/2016, str. 25–31, DOI:10.15199/13.2016.5.5
[4] Cook J. Farewell, Farecast: 2014. Microsoft kills airfare price predictor, to the dismay of its creator. *GeekWire*, April 2014.
[5] Mantin B., Gillen D.: 2011. The hidden information content of price movements. *European Journal of Operational Research*.
[6] Puller S.L., Taylor L.M.: 2012. Price discrimination by day-of-week of purchase: Evidence from the U.S. airline industry. *Journal of Economic Behavior & Organization*.
[7] Etzioni O., Knoblock C.A., Tuchinda R., Yates A.: 2003. To Buy or Not to Buy: Mining Airfare Data to Minimize Ticket Purchase Price. *SIGKDD*.
[8] Duhigg C.: 2013. Siła nawyku: dlaczego robimy to, co robimy i jak można to zmienić w życiu i biznesie. PWN Sp.z o.o.
[9] Wilks Y.: 2008. Machine Translation: Its Scope and Limits. Hamburg, Springer.