

*Michał Gamrot**

Zastosowanie technologii *big data* w e-biznesie

1. Wstęp

Obserwowane w ostatnich latach upowszechnienie się rozwiązań e-commerce, wraz z rozwojem technologii informatycznych, przyczyniło się do gwałtownego wzrostu ilości informacji dotyczących aktywności gospodarczej w internecie. Rosnąca każdego dnia ilość danych doczekała się nawet własnej nazwy – *big data*, co z angielskiego rozumieć można potocznie jako „duże dane” lub „masowe zbiory danych” (Witkowski 2014: 9–10). I choć duże zbiory danych mogą dotyczyć wszystkich sfer życia i biznesu, ich podstawowym źródłem jest internet (Kohavi, Provost 2001: 5–10).

Wśród teoretyków i praktyków e-biznesu dominuje przekonanie o istnieniu znaczącej wartości dodanej danych dotyczących aktywności gospodarczej w internecie (Harford 2014; Wielki 2014). Nie dzieje się to bez przyczyny – posiadając informacje o zachowaniu klientów podczas dokonywania decyzji o zakupie, przedsiębiorcy są w stanie dynamicznie dostosowywać wyświetlane treści do oczekiwań i preferencji użytkowników strony internetowej (Kohavi 2001). Dane dotyczące wielkości sprzedaży mogą z kolei pomóc w optymalizacji decyzji dotyczących wielkości stanów magazynowych, obniżając jednocześnie koszty logistyki. Za pomocą algorytmów *big data* możliwe jest również wykrywanie podejrzanych zachowań, minimalizując jednocześnie ryzyko oszustwa przez nieuczciwych klientów (Lek 2001: 160–165).

Wykorzystanie masowych zbiorów danych stanowi jednak duże wyzwanie. Samo posiadanie dużej ilości danych nie jest bowiem tożsame z posiadaniem dużej ilości informacji potrzebnych w procesie podejmowania decyzji. Aby surowe dane (ang. *raw data*) zostały przekształcone w wartościowe informacje, muszą być najpierw odpowiednio przygotowane, umieszczone w odpowiednim, zapewniającym prywatność i bezpieczeństwo miejscu oraz poddane skomplikowanej analizie statystycznej. Wszystkie te kroki, choć nie różnią się znacząco od standardowego podejścia do analizy, zaczynają być wyjątkowo problematyczne

* Wydział Ekonomiczno-Socjologiczny, Uniwersytet Łódzki.

w przypadku dużych zbiorów danych. Sytuację komplikuje dodatkowo szybkie tempo prowadzenia biznesu w Internecie. Dynamicznie zmieniające się otoczenie wymusza szybkie podejmowanie decyzji, co z kolei wymaga analizy danych w czasie rzeczywistym.

Artykuł przedstawia technologię *big data* oraz przybliża jej wykorzystanie w e-biznesie. W części drugiej (*Istota technologii big data?*) przedstawiona została koncepcja masowych zbiorów danych oraz powody rosnącego zainteresowania tą tematyką. Trzecia część artykułu (*Wykorzystanie big data w e-biznesie*) przybliża możliwości analizy masowych zbiorów danych oraz przedstawia przykłady dotychczasowych zastosowań technologii *big data* w e-biznesie. Część czwarta (*Wyzwania w analizie dużych zbiorów danych*) przybliża wyzwania i ograniczenia analizy dużych zbiorów danych. Artykuł kończy się przedstawieniem wniosków i rekomendacji dla szerszego wykorzystania technologii *big data* w praktyce e-biznesu oraz sugeruje możliwe kierunki dalszych badań.

2. Istota technologii big data

O ile potoczne rozumienie terminu *big data*, wynikające z jego bezpośredniego tłumaczenia, jest relatywnie zrozumiałe, o tyle dokładne zdefiniowanie tej koncepcji przysparza już wielu problemów.

Słownikowa definicja tego terminu z *Oxford English Dictionary*, określa *big data* jako „dane o bardzo dużym rozmiarze, zazwyczaj znacząco utrudniającym ich analizę” (Press 2014). Podobnej definicji dostarcza Wikipedia – „szeroko pojęty termin dotyczący zbiorów danych tak dużych, że ich przetwarzanie za pomocą tradycyjnych narzędzi staje się trudne” (Wikipedia 2015). W definicjach tych dominuje podkreślenie trudności w analizie danych. Warto w tym miejscu zaznaczyć, że obydwie definicje nie przedstawiają mierzalnej granicy oddzielającej dużej wielkości danych od standardowych. Wydaje się to być zrozumiałe, jako że ze względu na postęp technologiczny granica ta cały czas się przesuwa¹.

Innej definicji dostarcza szeroko cytowana publikacja analityków Gartnera (Laney 2001: 1–4). W definicji tej *big data* opisywana jest na podstawie trzech charakterystyk: wielkości (ang. *volume*), szybkości (ang. *velocity*) i zróżnicowania (ang. *variety*)².

¹ Pewnym problemem przedstawionych powyżej definicje jest odniesienie się do „tradycyjnych narzędzi”. Termin ten jest niedokładny i może sugerować zarówno komercyjne narzędzia (np. Microsoft Excel) jak i języki programowania o otwartym kodzie źródłowym (np. R). Różnice pomiędzy wielkością danych jakie są w stanie obsłużyć te narzędzia są relatywnie duże, stąd też definicje korzystające z koncepcji „tradycyjnych narzędzi” do analizy danych są relatywnie mało precyzyjne.

² Ze względu na angielskie nazwy wymienionych charakterystyk (*volume*, *velocity*, *variety*) definicję tą zwykło nazywać „3V”.

Podobnie jak w przypadku słownikowych definicji, podstawowym kryterium określenia danego zbioru danych jako masowego jest jego rozmiar. Laney, odnosząc się bezpośrednio do charakterystyki rynku *e-commerce*, podkreśla, że o ile zwiększająca się ilość zmiennych jakie charakteryzują każdą transakcję może przynieść dodatkowe korzyści, rozrastanie się zbiorów danych prowadzi do zwiększonych kosztów przechowywania i zarządzania danymi. O ile w dobie coraz tańszych dysków twardych argument ten nie jest już tak znaczący, korzyści ze zwiększania ilości obserwacji spadają wraz z każdą nową obserwacją. I choć szczegółowe dane dotyczące transakcji kupna komputera w sklepie internetowym sprzed 15 lat mają mniejszą wartość poznawczą niż analogiczna informacja sprzed miesiąca, koszt ich przechowywania jest taki sam.

Według Laney'a tym co wyróżnia masowe zbiory danych jest również częstotliwość z jaką powstają nowe obserwacje. Ma to szczególne znaczenie w e-biznesie, gdzie danymi mogą być nie tylko informacje dotyczące samych transakcji (np. wartość i rodzaj zamówionych dóbr), ale i sam proces zakupu. Algorytmy stosowane np. w sklepach internetowych są w stanie rozumieć i zapisywać kolejne kliknięcia użytkowników na stronie internetowej. Dzięki danym w tej postaci możliwe jest więc np. określenie kolejności przeglądanych przez użytkownika produktów, czasu jaki spędzał na oglądaniu poszczególnych towarów, czy też reakcji na wyświetlane na stronie reklamy.

Ostatnią z charakterystyk dużych zbiorów danych jest ich zróżnicowanie. Podobnie jak w przypadku rozmiaru i częstotliwości pojawiania się danych, cechą tą jest w głównej mierze domeną danych pochodzących z Internetu i wynika z dużej różnorodności form danych dostępnych w sieci. W przeciwieństwie do informacji gromadzonych w standardowych bazach danych (głównie tekstu i liczb), dane pochodzące z Internetu mogą mieć wiele różnych postaci – obrazów, nagrań wideo, animacji czy dokumentów. Stanowi to duże wyzwanie dla standardowych metod analitycznych, które w większości powstały przed upowszechnieniem się Internetu, a co za tym idzie nie są przydatne w analizie części danych pochodzących z e-biznesu.

O ile popularyzacja technologii *big data* jest w głównej mierze dziełem rewolucji informatycznej, a w szczególności rozwoju rynku *e-commerce*, nie jest to jedyny decydujący o tym czynnik. Równie ważnym czynnikiem jest popularyzacja mediów społecznościowych (np. Facebooka, Twittera) oraz upowszechnienie się technologii mobilnych (w szczególności smartfonów). Dzięki mediom społecznościowym możliwe jest dokładniejsze zrozumienie preferencji użytkowników Internetu oraz zrozumienie sieci relacji jakie tworzą. Z kolei wykorzystanie technologii mobilnych umożliwia dostęp do nowych rodzajów danych, m.in. danych sensorycznych ze smartfonów (np. GPS).

3. Wykorzystanie technologii *big data* w e-biznesie

Problemy biznesowe, jakie można rozwiązać dzięki wykorzystaniu masowych zbiorów danych, ograniczone są jedynie dostępnością samych danych. Przy dzisiejszym stopniu rozwoju technologii *big data* w e-biznesie, do głównych zastosowań tej technologii zaliczyć można:

- rekomendacje bazujące na wcześniejszych zakupach,
- dynamiczne ustalanie cen dóbr i usług,
- prewencję nadużyć.

Jednym z pierwszych zastosowań technologii *big data* była budowa algorytmów służących do rekomendacji dóbr i usług na podstawie wcześniej ocenionych lub zakupionych produktów (Cho, Kim 2004: 233–246). Algorytmy takie (np. *collaborative filtering*), działając na podstawie danych dotyczących już zrealizowanych zamówień, są w stanie znaleźć przedmioty często kupowane razem lub kupowane w odstępach czasu przez tego samego użytkownika. Posiadając dodatkowe informacje dotyczące niezalogowanych użytkowników odwiedzających stronę internetową (np. sprzętu z którego korzysta dana osoba) lub zalogowanych powracających klientów (np. płci, miejsca zamieszkania, czy wykształcenia) możliwe jest zbudowanie modelu wyjaśniającego zależność poszczególnych zmiennych z faktem decyzji o kupnie danego przedmiotu. Następnie, system jest w stanie wygenerować listę przedmiotów jaką dana osoba jest najbardziej skłonna kupić i wyświetlić je w miejscu gwarantującym największe prawdopodobieństwo zobaczenia.

Choć pierwsze zastosowania tego typu algorytmów dotyczyły głównie sprzedaży towarów konsumpcyjnych (np. sprzedaż książek przez serwis Amazon.com), obecne zastosowania rekomendacji opartej o technologię *big data* wybiegają daleko poza sprzedaż fizycznych dóbr. Przykładem są m.in. serwisy muzyczne oferujące spersonalizowane listy piosenek na podstawie wcześniej odsłuchanych utworów (np. Spotify, Apple Music), internetowe wypożyczalnie filmów (np. Netflix), czy aplikacje rekomendujące nowe restauracje na podstawie wcześniejszych ocen (np. Foursquare). Przykładem zastosowania podobnych algorytmów w Polsce jest m.in. serwis Filmweb i oferowana przez niego usługa Gustopodobni (www.filmweb.pl/similar-users) polegająca na rekomendacji filmów na podstawie wcześniej obejrzanych i ocenionych tytułów oraz rekomendacji innych użytkowników serwisu.

Innym zastosowaniem technologii *big data* w e-biznesie jest dynamiczne ustalanie cen w odpowiedzi na zmieniające się warunki rynkowe (Narahari i in. 2005: 231–256). Choć dynamiczne zmiany cen można zastosować we wszystkich branżach, ze względu na niższe koszty zmian cen w internecie (tzw. koszty menu), cena oferowanych w nim dóbr i usług może zmieniać się dynamicznie pod wpływem zmieniających się uwarunkowań gospodarczych (Chakrabarti, Scholnick 2005: 187–191). Dzięki możliwości dynamicznego dostosowania cen, moż-

liwe jest zwiększenie zysku poprzez bardziej wydajne zastosowanie dyskryminacji cenowej trzeciego stopnia – tj. sprzedawania tych samych (lub nieznacznie zmodyfikowanych) towarów po innych cenach różnym grupom konsumentów. W celu oszacowania ceny rezerwacji (tj. maksymalnej ceny po której dana osoba jest w skłonności do zakupu danego towaru), przydatne okazują się dane dotyczące sprzętu z którego korzysta przeglądający stronę internetową (np. systemu operacyjnego lub przeglądarki internetowej). Przykładem takiego zachowania jest Capital One, amerykański bank który w zależności od używanej przez klientów przeglądarki internetowej oferował różne oprocentowanie kart kredytowych (Lowrey 2010). Choć taktyka taka może przynieść zwiększone zyski w krótkim okresie, gdy dowiedzą się o niej klienci, może ona skutkować negatywnymi długookresowymi konsekwencjami. Przykładem mogą być negatywne reakcje konsumentów na informacje o dynamicznej manipulacji cenami filmów DVD sprzedawanych przez Amazon.com w 2005 roku (Weiss, Mehrotra 2001). Przypadek ten spowodował wzrost zainteresowania tematem dynamicznego ustalania cen przez organizacje konsumenckie (Turow i in. 2005), co w dużej mierze ograniczyło skalę wykorzystania tych technik w opisanym powyżej formie. Obecnie w celu skłonienia konsumentów do podjęcia decyzji o zakupie, algorytmy *big data* wykorzystywane są m.in. do identyfikacji klientów wahających się nad zakupem, tak aby móc nakłonić ich do zakupu (np. przez zaproponowanie dodatkowego rabatu) (Brown i in. 2011: 24–35).

Zastosowanie algorytmów dynamicznego dostosowania cen może być również wykorzystane przy sprzedaży dóbr i usług ograniczonych podażowo, jak np. bilety do kina czy loty samolotem. Przykładem tego typu usługi jest serwis Expedia (www.expedia.com), który umożliwia zakup biletów lotniczych po cenach niższych od standardowych. Ceny oferowanych biletów zmieniają się dynamicznie i zależą od popytu (szacowanego na podstawie ilości osób przeglądających dane połączenia) i podaży (mierzonej ilości wolnych miejsc w samolocie).

Ostatnim z omawianych zastosowań technologii *big data* w e-biznesie jest zapobieganie nadużyciom w handlu internetowym (Lek 2001: 160–165). Analiza dużych zbiorów danych może pomóc w identyfikacji podejrzanych zachowań, w szczególności oszustw finansowych polegających na dokonywaniu płatności za pomocą skradzionych kart kredytowych. W identyfikacji tych zachowań pomocne są dane dotyczące zarówno wcześniejszych transakcji, jak i informacje dotyczące sprzętu jakim posługuje się w konkretnej chwili dana osoba. Przykładem tego typu analizy może być analiza średniej wielkości zakupów danego użytkownika w zależności od np. pory dnia i urządzenia, z którego loguje się na stronie. W przypadku, w którym algorytm wykrywa nietypowe zachowanie (np. bardzo duża transakcja za pomocą nowego telefonu komórkowego, znajdującego się w chwili zakupu poza granicami kraju, z którego pochodzi dany klient), płatność zostaje wstrzymana a użytkownik musi dokonać dodatkowej weryfikacji, na

przykład potwierdzając chęć zakupu telefonicznie. W analizie tej kluczowe jest takie dostosowanie parametrów algorytmu, aby nie był on ani zbyt „restrykcyjny” (nie generował dużej ilości fałszywych alarmów), ani zbyt „liberalny” (nie ostrzegał przed możliwością popełnienia oszustwa zbyt rzadko). Sam dobór parametrów modelu jest więc w dużej mierze sztuką i wykracza poza standardowy zakres analizy *big data*.

4. Wyzwania w analizie dużych zbiorów danych

Wykorzystanie masowych zbiorów danych w e-biznesie, oprócz wspomnianych powyżej korzyści, ma również swoje ograniczenia.

Do podstawowych wyzwań, stojących na drodze szerszego wykorzystania technologii *big data*, jest wysoki koszt. Koszt ten w dużej mierze wynika z samego rozmiaru danych, które mogą być analizowane. Już samo przechowywanie danych wymaga poniesienia znacznych kosztów, związanych zarówno z infrastrukturą, jak i z zapewnieniem prywatności i bezpieczeństwa informacji (Greenberg 2008: 68–73). Przechowywanie dużych zbiorów danych wymaga nie tylko poniesienia większych kosztów, ale także zastosowania innej technologii. Masowe zbiory danych są tak duże, że niemożliwe jest utrzymanie ich na pojedynczym serwerze. Konieczne staje się więc zastosowanie technologii rozproszonych takich jak np. Apache Hadoop (hadoop.apache.org), dzięki którym dane mogą być przechowywane na wielu serwerach jednocześnie, często również w różnych lokalizacjach.

Kolejnym kosztem jest transformacja danych. Choć dane pochodzące z Internetu są relatywnie wiarygodne i pozbawione manualnych błędów (w porównaniu np. do danych zbieranych przez ankietatorów), praktyka analityczna pokazuje (Osborne 2012), że zanim zostaną przeanalizowane, dane te muszą zostać poddane procesowi czyszczenia (ang. *data cleaning*). Ma to szczególne znaczenie w przypadku wykorzystania danych z różnych źródeł – np. systemu sprzedaży, sieci społecznościowych, mediów oraz stron internetowych konkurencji. Agregacja danych z tak wielu źródeł, choć bardzo pomocna w późniejszej analizie, jest niezwykle trudna. Ze względu na brak powszechnych standardów pomagających w identyfikacji dóbr sprzedawanych w Internecie³, trudno jest określić automatycznie czy dane dobro jest tym samym co np. oferowane przez konkurencję lub wspomniane przez użytkowników mediów społecznościowych. Z tych powodów czyszczenie danych jest jedną z najbardziej pracochłonnych czynności (Rahm, Do 2000: 3–13), będąc jednocześnie jedną z największych przeszkód przed szerszym wykorzystaniem technologii *big data* w praktyce.

³ Przykładem takiego standardu w przypadku dokumentów może być *Digital Object Identifier* (DOI), w przypadku książek *International Standard Book Number* (ISBN).

Masowe zbiory danych stanowią również poważne wyzwanie analityczne. Ilość zmiennych i obserwacji jest często tak duża, że problematyczne może okazać się nawet obliczenie podstawowych statystyk opisowych, nie wspominając o bardziej wysublimowanych modelach ekonometrycznych. Analiza dużych zbiorów danych wymaga więc wykorzystania specjalistycznych algorytmów uczenia maszynowego (np. klastrowania czy drzew decyzyjnych) lub specjalnie przystosowanych do tego wersji znanych metod analitycznych (np. modeli regresji liniowej przystosowanej do dużych zbiorów danych). I choć istnieją darmowe, oparte o otwarty kod oprogramowania, narzędzia do analizy *big data*, konieczność zastosowania specjalistycznych narzędzi jest głównym determinantem wysokiego kosztu praktycznego wykorzystania dużych zbiorów danych. Wynika to bezpośrednio z nadal niewielkiej podaży specjalistów z zakresu zaawansowanej analizy danych (ang. *data science*), co przy wysokim popycie na ich usługi, czyni ich pracę bardzo drogą (King, Magoulas 2014).

Nie wszystkie ograniczenia stojące za szerszym zastosowaniem technologii *big data* są związane z czynnikami ekonomicznymi. Równie ważne, o ile nie ważniejsze, są argumenty natury metodologicznej. O ile bowiem wysokie koszty wdrożenia masowych zbiorów danych można uznać za tymczasowe (i wynikające z wczesnej fazy rozwoju tej technologii), problemy metodologiczne są niezależne od posiadanych przez nas zasobów finansowych.

Podstawowym, nieekonomicznym ograniczeniem w wykorzystaniu masowych zbiorów danych jest przekonanie o tym, że większa ilość danych prowadzi *zawsze* do dodatkowych wniosków (Harford 2014). O ile analiza na podstawie danych jest najczęściej lepsza od analizy na podstawie przeczuć i zasłyszanych anegdot, powiększanie zbioru danych przeznaczonego do analizy o kolejne zmienne i obserwacje przynosi zazwyczaj coraz mniejsze efekty. Wartość dodana każdej dodatkowej zmiennej opisującej dane wydarzenie jest coraz mniejsza⁴. I choć z punktu widzenia teorii, im większa jest ilość dostępnych danych, tym większa precyzja estymacji, z punktu widzenia praktyki biznesowej, zebranie, przechowanie, transformacja i analiza każdej dodatkowej zmiennej obciążona jest dodatkowymi kosztami. Można więc stwierdzić, że w praktyce istnieje optimum pomiędzy ilością danych a wartością dodaną dla przedsiębiorstwa.

Z drugiej strony, nawet w e-biznesie nie możemy mieć pewności, że dysponujemy wszystkimi danymi. Choć w przypadku działalności w internecie, gdzie każda interakcja pozostawia za sobą ślad w postaci dodatkowych danych, twierdzenie o niewystarczającej ilości danych wydaje się dalece kontrowersyjne, w rzeczywistości, dane pochodzące jedynie z Internetu nie mogą być uznane za

⁴ W skrajnych przypadkach zwiększenie zbioru danych o kolejne obserwacje może przynieść efekty odwrotne do zamierzonych. Ze względu na to, że analiza dużych danych jest skomplikowana obliczeniowo, uwzględnienie coraz to większych zbiorów danych może wymuszać uproszczenie postaci funkcyjnej modelu co w konsekwencji może prowadzić do spadku jakości predykcji.

wierne odwzorowanie rzeczywistości gospodarczej. Koncentrując się wyłącznie na danych pochodzących z sieci, choć jesteśmy w stanie dokładnie zrozumieć zachowanie osób odwiedzających *naszą* stronę internetową, możemy ulec iluzji homogeniczności całej populacji i uogólnić obserwowane w sieci zachowania na osoby niekorzystające z naszych usług (Harford 2014: 14–19).

Ostatnim ograniczeniem zastosowania technologii *big data*, z punktu widzenia użytkowników Internetu, jest obawa o wpływ, jaki zbieranie dużych ilości danych na ich temat ma na ich prywatność (Tene, Polonetsky 2012: 63–69). W ocenie części użytkowników Internetu, *nadmierna* ilość informacji, jakie wiedzą na ich temat korporacje, może prowadzić do powstania dysproporcji w poziomie wiedzy dotyczącej bodźców skłaniających ich do podejmowania decyzji o zakupach. To z kolei może sprawiać, że będą oni manipulowani w celu kupowania coraz to większej ilości niepotrzebnych im dóbr. Dalszy sukces technologii *big data* w dużej mierze zależeć będzie więc od zmian w percepcji konsekwencji udostępniania danych przez użytkowników stron *e-commerce*.

5. Podsumowanie

Wykorzystanie technologii *big data* ma swoje ograniczenia. Wczesna faza rozwoju tej technologii sprawia, że wykorzystywane w niej narzędzia nie są jeszcze powszechnie znane – a co za tym idzie koszt zatrudnienia umiających się nimi posługiwać specjalistów jest relatywnie wysoki. Wysoki jest również koszt przechowywania, transformacji i analizy dużych zbiorów danych, co przy całym czasie relatywnie niewielkiej wiedzy na temat potencjalnych zysków z zastosowania tej technologii, przyczynia się do nadal niewielkiego wykorzystania jej w praktyce.

Pomimo wspomnianych ograniczeń, analiza coraz większych zbiorów danych może pomóc praktykom e-biznesu w lepszym zrozumieniu prowadzonego przez siebie przedsiębiorstwa. Dalszy sukces technologii *big data* w transformacji e-biznesu zależeć będzie od kilku czynników:

- popularyzacji badań dotyczących efektów zastosowań technologii *big data* w praktyce,
- spadku kosztów analizy,
- zmian w percepcji konsekwencji udostępniania danych przez użytkowników.

Jeśli uda się zrealizować powyższe cele, technologia *big data* stanie się jednym z głównych motorów zmian e-biznesu a szersze wykorzystanie informacji z mediów społecznościowych, aplikacji mobilnych i danych dotyczących kliknięć na stronach internetowych, umożliwi dokładniejsze zrozumienie potrzeb klientów, optymalizację treści marketingowych czy zmniejszenie ryzyka popełnienia nadużyć.

BIBLIOGRAFIA

- Brown B., Chui M., Manyika J. (2011), *Are you ready for the era of 'big data'*, „McKinsey Quarterly”, vol. 4, p. 24–35.
- Chakrabarti R., Scholnick B. (2005), *Nominal rigidities without literal menu costs: evidence from E-commerce*, „Economics Letters”, vol. 86(2), p. 187–191.
- Cho Y. H., Kim J. K. (2004), *Application of Web usage mining and product taxonomy to collaborative recommendations in e-commerce*, „Expert systems with Applications”, vol. 26(2), p. 233–246.
- Greenberg A., Hamilton J., Maltz D. A., Patel P. (2008), *The cost of a cloud: research problems in data center networks*, ACM SIGCOMM Computer Communication Review, vol. 39(1), p. 68–73.
- Harford T. (2014), *Big data: A big mistake?*, „Significance”, vol. 11(5), p. 14–19.
- Harford T. (2014), *Big data: Are we making a big mistake?*, <http://goo.gl/rKNqtg>, (data dostępu 21.06.2015).
- King J., Magoulas R. (2014), *2014 Data Science Salary Survey. Tools, Trends, What Pays (and What Doesn't) for Data Professionals*, O'Reilly.
- Kohavi R. (2001), *Mining e-commerce data: the good, the bad, and the ugly*, [w:] *In Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, p. 8–13.
- Kohavi R., Provost F. (2001), *Applications of data mining to electronic commerce*, Springer US, p. 5–10.
- Laney D. (2001), *3D data management: Controlling data volume, velocity and variety*, „META Group Research Note”, vol. 6, p.1–4.
- Lek M., Anadarajah B., Cerpa N., Jamieson R. (2001), *Data Mining Prototype for Detecting eCommerce Fraud*, [w:] *European Conference on Information Systems Proceedings*, vol. 60, p. 160–165.
- Lowrey A. (2010), *How much is that doggie in the browser window?*, „Slate”, <http://goo.gl/xDxdbp>, (data dostępu 21.06.2015).
- Narahari Y., Raju C. V. L., Ravikumar K., Shah S. (2005), *Dynamic pricing models for electronic business*, „In Sadhana” (Academy Proceedings in Engineering Sciences), vol. 30(2–3), p. 231–256.
- Osborne J. W. (2012), *Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data*. Sage Publications.
- Press G. (2014), *12 Big Data Definitions: What's Yours?*, „Forbes”, <http://goo.gl/2m9Xie>, (data dostępu 21.06.2015).
- Rahm E., Do H. H. (2000), *Data cleaning: Problems and current approaches*, „IEEE Data Engineering Bulletin Issues”, vol. 23(4), p. 3–13.
- Tene O., Polonetsky J. (2012), *Privacy in the age of big data: a time for big decisions*, „Stanford Law Review Online”, vol. 64, p. 63–69.
- Turow J., Feldman L., Meltzer K. (2005), *Open to Exploitation: America's Shoppers Online and Offline*, A Report from the Annenberg Public Policy Center of the University of Pennsylvania, http://repository.upenn.edu/asc_papers/35, (data dostępu 21.06.2015).
- Weiss R. M., Mehrotra A. K. (2001), *Online dynamic pricing: efficiency, equity and the future of e-commerce*, „Virginia Journal of Law and Technology”, vol. 6(1).
- Wielki J. (2014), *Analiza możliwości wykorzystania zjawiska Big Data w e-biznesie*, Prace Naukowe Uniwersytet Ekonomiczny w Katowicach, Katowice.
- Witkowski J. (2014), *Statystyka oficjalna wobec wyzwań globalnych*, „Wiadomości Statystyczne”, nr 4(635), s. 9–10.

THE USE OF BIG DATA IN E-COMMERCE

SUMMARY

Big data, defined as a process of storing, transforming and analyzing large volumes of data, is one of the most discussed issues in e-commerce. By gathering large quantities of data about their customers, e-commerce practitioners can gain valuable insights regarding (1) the reasons of why people are making purchases and (2) maximum prices they are willing to pay for certain goods, in order to maximize profit margins. By analyzing customer's behavior it's also possible to (3) identify patterns that indicate fraud, which in turn might greatly reduce costs and risks of e-commerce business.

Big data also has its limitations. Storing, transforming and analyzing large quantities of data is costly and because big data is still a relatively new technology, the supply of skilled professionals is low. Big data poses also some difficult methodological concerns, i.e. whether or not more data always leads to better insights and whether online data shouldn't be supplemented by external, more traditional sources. The paper mentions also the issue of privacy, observing that it's one of the main factors affecting success of big data technologies in the future.