

Janusz JANCZYK

*Dr inż., Firma edukacyjna EDU-ICT, ul. Wrocławska 11, 41-200 Sosnowiec;
janusz.janczyk@gmail.com*

BIG DATA W RELACJI DO PROCESÓW ZMIAN W EDUKACJI

BIG DATA IN RELATION TO THE PROCESSES OF CHANGE IN EDUCATION

Słowa kluczowe: Big Data, edukacja, rozwój.
Keywords: Big Data, education, growth.

Streszczenie

Implementacje technologii informacyjnych w sferę edukacji cechują się od wielu lat wysoką dynamiką wzrostu. Najbardziej intrygującym i nierozpoznanym obecnie zjawiskiem, przenikającym nie tylko edukację, ale wszelkie działania ludzkie jest Big Data. Należy skorzystać z nowych możliwości IT w takim zakresie, aby przyspieszyć rozwój edukacyjnych zastosowań Big Data i ograniczyć wpływ dostrzeganych oraz przyszłych zagrożeń. Podjęto próbę ewaluacji technologii Big Data w obszarze indeksowania, pozyskiwania i dystrybucji informacji w kontekście usieciowionej edukacji.

Summary

Implementation of information technologies in the sphere of education are characterized by many years of high growth. The most intriguing and currently unrecognized phenomenon, permeating not only education, but all human activities is Big Data. Should take advantage of new IT capabilities in extent to accelerate the development of educational applications of Big Data and limit the impact of perceived and future threats. An attempt was made evaluation of Big Data technology in the field of indexing, acquisition and distribution of information in the context of networked education.

Wstęp

Zjawisko Big Data nie jest niczym nowym i od lat charakteryzuje się olbrzymim wzrostem. Badania, obserwacje, dyskusje międzynarodowe i inicjatywy wielu rządów wydają się być ukierunkowane na wykorzystanie potencjału technologii pozyskiwania i analizowania ogromnych ilości danych, przy zasto-

sowaniu coraz bardziej zaawansowanych metod informatycznych. Ludzie pracujący w obszarze działalności gospodarczej od wielu lat zbierają i łączą olbrzymie zestawy danych w celu poprawy segmentacji klientów i towarów, aby lepiej zrozumieć reguły funkcjonowania rynków gospodarczych¹. W dziedzinie edukacji Big Data nadal prezentuje stosunkowo niszowe zastosowania, pomimo że są one szeroko udokumentowane i są upowszechnione narzędzia IT (technologii informacyjnych).

Big Data – zarys problematyki i cechy określające

W związku z pojawieniem się nowych technologii, urządzeń i środków komunikacji (np. portale społecznościowe dostępne mobilnie) ilość danych wytwarzanych przez ludzkość rośnie w tempie zatrważającym specjalistów IT. Wielość danych wytworzonych przez ludzi od początku czasu do 2003 roku jest szacowana na ok. 5 mld gigabajtów. Jeśli przechowuje się dane zapisując je na dyskach, to można tymi dyskami wypełnić całe boisko do piłki nożnej. Taką samą ilość danych tworzono co każde dwa dni w 2011 roku oraz w każde dziesięć minut w roku 2013. Wskaźnik ten wciąż ogromnie rośnie, na co dowodem jest to, że 90% danych na świecie wygenerowanych zostało w ciągu ostatnich kilku lat². Ten ogromny boom danych doczekał się określenia – „Big Data”.

Czym właściwie jest Big Data?

Oznacza niewyobrażalnie dużą ilość danych i jest to przestrzeń takich ogromnych zbiorów danych, że nie sposób je przetwarzać przy użyciu tradycyjnych technik i metod informatyki. Big data to nie tylko same dane, raczej coś znacznie obszerniejszego, co obejmuje różne: narzędzia, techniki i struktury. Termin „Big Data” może być też nieco mylący, gdyż oznaczałoby to, że wcześniej zarejestrowane dane były jakiegoś małe i że jedynym wyznacznikiem jest sama wielkość? Otóż Big Data stosuje się do danych, które nie mogą być przetwarzane lub analizowane przy użyciu tradycyjnych metod, procesów i narzędzi. Coraz częściej różnego rodzaju organizacje stoją przed wyzwaniem, określanym mianem Big Data. Posiadają dostęp do bogactwa, ogromu informacji, lecz nie wiedzą jaka jest jej wartość, gdyż jest ona dostępna w najbardziej „surowej” postaci. Już w pierwszej dekadzie XXI wieku w zasobach edukacyjnych Internetu obserwować można było tego typu zjawisko. Pomimo że nie wiązano tego z Big Data, to wielość, a zwłaszcza różnorodność formatów z treściami edukacyjnymi, publikowanymi w Internecie, stwarzała problemy z ich klasyfikacją,

¹ Por. J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. Hung Byers, *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, McKinsey report: McKinsey Global Institute 2011.

² Por. V. Mayer-Schonberger, K. Cukier, *Big Data. Rewolucja, która zmieni nasze myślenie, pracę i życie*, Wyd. MT Biznes, Warszawa 2014.

indeksowaniem, a szczególnie ewaluacją. Dostępne dane o niewątpliwej wartości edukacyjnej nie były usystematyzowane, gdyż nie poddawały się procesom strukturyzacji. Nie mogły być traktowane jako Big Data, ponieważ ich ilość nie koresponduje w żaden sposób z wielkością opisywanego zjawiska.

Problematyka Big Data jest w pełni obowiązująca od kilku lat, gdyż świat od dawna trwa w ciągłej zmianie – paradygmat społeczeństwa informacyjnego. Poprzez stosowane mechanizmy, jesteśmy w stanie odkryć więcej, a jeśli tak, to stwarzana jest tendencja do dalszego testowania i przechowywania danych (lub przynajmniej niektórych z nich). Przez postęp w ICT, ludzie i rzeczy stają się bardziej połączeni, i nie tylko zmiennie w czasie, ale ciągle. Takie połączenie stwarza efekt tzw. uciekającego pociągu (ang. *runaway train*), gdyż ludzie nie nadążają za połączeniami urządzeń ICT. Łączność pośrednia, ogólnie określana jako maszyna-maszyna (M2M), jest odpowiedzialna za dwucyfrowe tempo wzrostu rok do roku. W związku z tym, że układy scalone są obecnie bardzo tanie, możliwe jest zwiększanie „inteligencji” wszystkich wytwarzanych urządzeń ICT. W tym kontekście zrozumiałe jest generowanie 61,5% ruchu danych w Internecie przez boty³.

Według specjalistów z IBM⁴ trzy cechy definiują Big Data: *ilość*, *różnorodność* i *prędkość*. Zainicjowali oni także potrzebę nowej klasy umiejętności do poszerzania dróg myślenia o Big Data, aby zapewnić lepszą ścieżkę rozwoju i kontrolę nad istniejącymi dziedzinami wiedzy i zdolnościami do działania w ich obszarach.

- *Ilość danych*

Występuje poważna obawa, że ogromna ilość danych, przechowywanych w zasobach (magazynach) sieci rozległej w końcu eksploduje. W 2000 roku przechowywanych było ok. 800 000 PB (petabajtów). Wypada wspomnieć, że już obecnie wiele tworzonych danych nie podlega analizie i nie ma szans na jakąkolwiek analizę w przyszłości. Według badaczy z IBM⁵ należy się spodziewać wzrostu liczby danych do 35 ZB (zettabajtów) w 2020 roku. W serwisie Twittera generuje się dziennie ponad 7 TB danych, w serwisie Facebook 10 TB dziennie, a są przedsiębiorstwa generujące co godzinę terabajty danych – każdego dnia w roku. Jak wykazuje problematyka Big Data, organizacje wciąż borykają się z ogromnymi ilościami danych. Organizacje, które nie wiedzą, jak zarządzać tymi danymi są przytłoczone, pomimo istniejących możliwości stosowania nowoczesnych platform technologicznych. Potrafią one analizować prawie wszyst-

³ P. Kreft, *Większość ruchu w internecie generują... boty*, serwis Komputer Świat, <http://www.komputerswiat.pl/nawosci/internet/2013/50/wiekszosc-ruchu-w-internecie-generuja-boty.aspx> (data publikacji: 14.12.2013).

⁴ P.C. Zikopoulos, C. Eaton, D. deRoos, T. Deutsch, G. Lapis, *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*, McGraw-Hill books, USA 2012.

⁵ *Ibidem*.

kie dane lub przynajmniej te, które zostaną uznane za istotne. Ma to na celu lepsze zrozumienie własnej działalności, klientów i całego rynku. Ponieważ ilość dostępnych danych dla przedsiębiorstw wzrasta, pewien procent danych można przetwarzać, analizować i zrozumieć, aby zmniejszyć strefę danych nierozpoznanych. Co znajduje się w tej strefie danych nierozpoznanych? Analitycy z IBM przekonują⁶, że trudno określić, czy jest to coś ważnego lub może w ogóle nic, ale ta niewiedza stwarza problem główny. Problematyka ilości danych, na przestrzeni ostatnich lat, zmienia się z terabajtów do petabajtów z nieuchronną zmianą do zettabajtów, a wszystkie te dane z różnych powodów nie mogą być przechowywane w swoich tradycyjnych systemach.

▪ *Różnorodność danych*

Objętość danych związana ze zjawiskiem Big Data niesie ze sobą inne, nowe wyzwania dla centrów danych – trzeba sprostać tzw. różnorodności (ang. *variety*). Z eksplozji czujników i inteligentnych urządzeń, a także technologii współpracy społecznej, dane w przedsiębiorstwie stają się niekomunikatywne, ponieważ obejmują nie tylko tradycyjne relacyjne dane, ale także surowe, semi-strukturalne i niestructuralne dane ze stron internetowych, plików dzienników internetowych (w tym dane click-stream), indeksów wyszukiwania, forów społecznościowych, e-maili, dokumentów, dane z czujników systemów aktywnych i pasywnych itd. Ponadto, tradycyjne systemy mogą usiłować przechowywać i wykonywać wymagane analizy, aby wydobyć informacje z treści tych dzienników, ponieważ większość informacji generowanych nie nadaje się do tradycyjnych technologii bazodanowych. Według doświadczonych specjalistów IBM⁷, niektóre firmy idą tą ścieżką i w większości zaczynają rozumieć możliwości Big Data. Aby wykorzystać możliwości Big Data, przedsiębiorstwa muszą być w stanie przeanalizować wszystkie typy danych, zarówno relacyjne i nierelacyjne: tekst, dane z czujników, audio, wideo, dane transakcyjne i inne.

▪ *Prędkość danych*

Jak w wielości i różnorodności gromadzonych oraz przechowywanych danych zachodzą zmiany, tak też stąd dane mają prędkość, z którą są generowane i muszą być obsługiwane. Konwencjonalne rozumienie prędkości zazwyczaj polega na tym, jak szybko dane są przenoszone i przechowywane, co jest związane ze wskaźnikami pobierania. Dobrze jest szybko zarządzać Big Data z powodu ilości danych, co jest naturalną konsekwencją szybkości docierania danych do celu. Ten kontekst prędkości jest o wiele bardziej interesujący, niż w przypadku tradycyjnych definicji.

Właściwe podejście do myślenia o problemie prędkości danych należy rozpocząć w momencie generowania danych. Nie należy ograniczać identyfikacji

⁶ *Ibidem.*

⁷ *Ibidem.*

prędkości do stóp wzrostu powiązanych z repozytoriami danych, zaleca się stosowanie definicji do tzw. danych w ruchu – szybkość, z jaką dane przepływają. W takim ujęciu specjaliści przyjmują, że dzisiejsze przedsiębiorstwa mają do czynienia z petabajtami danych zamiast terabajtów, a wzrost zastosowań czujników RFID i innych strumieni informacyjnych doprowadził do stałego przepływu danych w takim tempie, które uniemożliwia stosowanie tradycyjnych systemów obsługi. Coraz więcej danych wytwarzanych obecnie ma bardzo krótki okres przechowywania, więc organizacje muszą analizować te dane w czasie rzeczywistym, gdy tylko będą miały do nich dostęp. Przetwarzanie strumieni w skali Big Data to koncepcja zaproponowana przez specjalistów z IBM⁸, która jest wdrażana od kilku lat i służy jako nowy paradygmat dla problemu Big Data. W tradycyjnej obróbce danych można tylko uruchamiać zapytania tylko z danymi statycznymi. Z przetwarzania strumieni danych można utworzyć proces podobny do ciągłego zapytania (dynamicznych zapytań), który identyfikuje dane już w czasie ich powstawania.

Skuteczność radzenia sobie z problemem Big Data wymaga wykonania analiz wobec wielkości i różnorodności danych, gdy są one jeszcze w ruchu, a nie gdy dotrą do celu. Najlepsze rozwiązanie dla platformy obsługi prędkości Big Data definiuje się, jako zmienną integracji z wyłączeniem partii danych do wglądu (w koncepcji Apache Hadoop) lub do wglądu partii danych w połączeniu z tzw. strumieniem-danych-na-drucie (koncepcja IBM). Takie podejście prezentuje spójną platformę Big Data, którą można wykorzystać w obu obszarach analizy danych (statycznym i dynamicznym). Jednakże właściwe jest podjęcie analizy w czasie rzeczywistym transmisji strumieniowej, z wglądem do inicjacji badań strumieni w oparciu o pojawiające się nowe dane⁹.

Technologie stosowane w Big Data

Technologie stosowane w obszarze Big Data są ważne dla zapewnienia dokładnej analizy danych, co prowadzi do pełnego skonkretyzowania podejmowanych decyzji, w wyniku większej wydajności operacyjnej, obniżeniu kosztów i zmniejszeniu ryzyka dla biznesu. Powstało w związku z tym pytanie: jakiej infrastruktury należy spodziewać się, aby wykorzystać zasoby Big Data? Takie, którymi można zarządzać i przetwarzać w nich ogromne ilości danych strukturalnych i niestructuralnych, w dodatku w czasie rzeczywistym. Ponadto te struktury powinny chronić prywatność i zapewniać bezpieczeństwo użytkowanych danych¹⁰.

⁸ *Ibidem.*

⁹ *Ibidem.*

¹⁰ Por. P. Aleksandrowicz, *Wielkie dane czy wielki brat*, <https://www.nbportal.pl/wiedza/recentje/wielkie-dane-czy-wielki-brat> (dostęp: 23.10.2014 r.).

Istnieją różne technologie na rynku do obsługi Big Data, od różnych dostawców, w tym Amazon, IBM, Microsoft. Przyglądając się technologiom, które obsługują duże ilości danych, wyróżnia się dwie najbardziej typowe¹¹:

- *Operacyjna Big Data*

Obejmuje ona systemy, które zapewniają – jak MongoDB, możliwości operacyjne w czasie rzeczywistym, interaktywne obciążanie, gdzie dane są przede wszystkim przechwytywane i zapisywane.

Systemy NoSQL Big Data są zaprojektowane tak, aby skorzystać z nowych architektur *cloud computing*, które pojawiły się w ciągu ostatniej dekady, aby umożliwić tanie i skuteczne uruchomienia masowych obliczeń. To sprawia, że dużymi obciążeniami danych, operacyjnie znacznie łatwiej się zarządza, a także jest to tańsze i szybsze w realizacji.

- *Analityczna Big Data*

Obejmuje ona systemy takie, jak tworzenie Procesów Masowo Równoległych (MPP) dla systemów baz danych i MapReduce, które zapewniają możliwości analityczne dla retrospektywnej i kompleksowej analizy, a zatem mogą wykonać analizę większości lub wszystkich danych. MapReduce zapewnia nową metodę analizy danych, która jest komplementarna do możliwości oferowanych przez SQL. Dodatkowo systemy oparte na MapReduce są skalowalne w górę od pojedynczych serwerów do tysięcy maszyn o różnych mocach obliczeniowych.

Te dwie klasy technologii są komplementarne i często stosowane razem.

Relacje edukacji z Big Data

Dla niektórych, Big Data reprezentuje zmianę paradygmatu o sposobie zrozumienia i odkrywania otaczającego świata, a przynajmniej, że jest postrzegana, jako sposób na lepsze wykorzystanie i twórcze analizowanie częściowych danych, najlepiej użytecznych dla pożytku publicznego i prywatnego. Obszary edukacyjnych eksploracji danych i uczenia analitycznego są opracowywane zarówno dla charakterystyk w różnych dziedzinach nauczania¹², jak i dla rządów są przygotowywane sprawozdania na temat potencjału Big Data w edukacji¹³.

¹¹ Por. T.H. Davenport, J. Dyché, *Big Data in Big Companies*, International Institute of Analytics, Report May 2013.

¹² Por. R. Ferguson, *The State Of Learning Analytics in 2012: A Review and Future Challenges*, Technical Report KMI-12-01, Knowledge Media Institute, The Open University, UK 2012, <http://kmi.open.ac.uk/publications/techreport/kmi-12-01> (dostęp: 21.01.2013 r.), a także por. C. Romero, S. Ventura, *Educational Data Mining: A Review of the State of the Art*, IEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews 40(6)/2010.

¹³ Por. M. Bakia, L. Shear, Y. Toyama, A. Lassetter, *Understanding the Implications of Online Learning for Educational Productivity*, Washington, D.C., 2012, <http://tech.ed.gov/files/2013/10/implications-online-learning.pdf> (dostęp: 15.10.2013 r.).

Jednak chyba największym zaskoczeniem jest to, że dyskurs wokół Big Data i edukacji przenika problematyka efektywności i opłacalności, zarówno w zakresie wykorzystania danych do celów poprawy efektów kształcenia – „dostarczania wiedzy” oraz jako sposób prowadzenia badań w tej dziedzinie. Ważne jest, aby dyskusja nad wykorzystaniem Big Data w celu zwiększenia wydajności, przejrzystości, wspierania konkurencyjności, a jako narzędzie do oceny efektywności (szkół i nauczycieli) trafiła na pierwszy plan do debaty akademickiej¹⁴.

W czasach kryzysu, komercyjne dyskursy stają się bardziej istotne w debatach nt. edukacji. Tak więc, w pewnych kręgach Big Data jest coraz bardziej promowana jako forma „modyfikacji technicznej” dla badań edukacyjnych i zastosowań w procesach kształcenia¹⁵, pomijając użycie tych samych narzędzi do wzmocnienia pozycji i wsparcia edukacyjnych badań praktycznych i krytycznych. W edukacyjnej debacie, dotyczącej relacji Big Data i edukacji podnoszone są zwłaszcza trzy obszary¹⁶:

- obszar wyzwań dotyczący etyki, gdyż Big Data implikuje szereg rozważań etycznych, zwłaszcza wokół prywatności, świadomej zgody i ochrony danych osobowych, i podnosi najważniejsze pytania: jakiego rodzaju dane powinny być łączone, indeksowane i analizowane oraz jakie cele obligują do zbierania tych danych?;
- obszar wyzwań dotyczących zapewnienia zrozumienia rodzajów badań, które mogą lub nie mogą być prowadzone przy użyciu Big Data. Dostępność Big Data ogranicza rodzaje pytań, które można zadać, a możliwe jest analizowanie tylko tych danych, które dają się zbierać lub już są zarejestrowane. Nie ma danych dobrych na wszystko, więc dostępność Big Data oddaje obraz tego co ludzie badają, czego poszukują i o co pytają;
- obszar wyzwań skupiających się wokół kwestii nierówności. Big Data może zarówno wzmacniać, a może też generować nowe nierówności społeczne i edukacyjne w wielu aspektach. Najważniejszy z nich zasada się wokół kwestii wyboru strumieni danych do analizowania przy użyciu Big Data. Z dużym prawdopodobieństwem można stwierdzić, że dane tylko tych osób będą reprezentowane lepiej w badaniach, które częściej używają Internetu.

¹⁴ Por. R. Eynon, *The rise of Big Data: what does it mean for education, technology, and media research?*, Learning, Media and Technology, Volume 38, Issue 3, 2013, Published online: 25 Feb 2013, <http://www.tandfonline.com/doi/abs/10.1080/17439884.2013.771783> (dostęp: 25.02.2013).

¹⁵ Por. K. Robins, F. Webster, *The Technical Fix: Education, Computers and Industry*, Macmillan, New York 1989.

¹⁶ Por. R. Eynon, *The rise of Big Data...*

Zakończenie

Nie powinno się zakładać, że Big Data w swym ekscytującym rozwoju nie nadaje się do badań w obszarach wielu dziedzin nauki, mediów i technologii. Szczególnie w obszarze edukacji należy samemu kształtować implementacje Big Data, nie czekając na rozwiązania z innych dyscyplin¹⁷. Może się zdarzyć, że na pierwszy rzut oka wnioski z badań nad Big Data zaoferują wiele zgrabnych odpowiedzi, szczególnie w porównaniu do powierzchownego korzystania z bardziej uznanych metod badawczych nauk społecznych.

W badaniach własnych nad netykietą wśród polskich nauczycieli¹⁸ napotkano braki w spójności indeksacji informacji z różnych obszarów aktywności edukacyjnej. Problematyka Big Data jest wynikiem wcześniejszych technicznych i organizacyjnych kryzysów występujących w środowisku sieci rozległej¹⁹. W zakresie indeksowania i wyszukiwania informacji z obszaru edukacji (np. w badaniu dokumentów) z implementacjami Big Data nie wystarczy, prezentowane początkiem XXI wieku, racjonalne podejście do użytkowania zasobów Internetu²⁰. Rozwiązania technologiczne Big Data stwarzają możliwości dotarcia do informacji/wiedzy wydawałoby się, że bardziej wysublimowanej, ale także uwikłanej w inne, nieedukacyjne, konteksty.

Bibliografia

- Aleksandrowicz P., *Wielkie dane czy wielki brat*, Portal Edukacji Ekonomicznej, <https://www.nbportal.pl/wiedza/recenzje/wielkie-dane-czy-wielki-brat> (dostęp: 23.10.2014 r.).
- Bakia M., Shear L., Toyama Y., Lasseter A., *Understanding the Implications of Online Learning for Educational Productivity*, Washington D.C. 2012, <http://tech.ed.gov/files/2013/10/implications-online-learning.pdf> (dostęp: 15.10.2013 r.).
- Davenport T.H., Dyché J., *Big Data in Big Companies*, International Institute of Analytics, Report May 2013.
- Drachler H., Wolfgang G., *Confidence in Learning Analytics*, Conference: Learning Analytics and Knowledge, Canada, Vancouver 2012 (dostęp: 29.04.2012 r.).
- Eynon R., *The rise of Big Data: what does it mean for education, technology, and media research?*, Learning, Media and Technology, Vol. 38, Issue 3, 2013, <http://www.tandfonline.com/doi/abs/10.1080/17439884.2013.771783> (dostęp: 25.02.2013 r.).

¹⁷ Por. H. Drachler, G. Wolfgang, *Confidence in Learning Analytics*, Conference: Learning Analytics and Knowledge, Canada, Vancouver 2012 (dostęp: 29.04.2012 r.).

¹⁸ Por. J. Janczyk, *Netykieta w środowisku nauczycieli przedmiotów informatycznych* [w:] *Historia, instytucjonalizacja i perspektywy kształcenia nauczycieli na Śląsku – Jubileusz 80-lecia*, red. S. Juszczyk, D. Morańska, Wyd. Uniwersytet Śląski, Katowice 2010.

¹⁹ Por. J. Janczyk, *Technical and Organisational Crises in Nets*, TRANSFORMACJE Special Issue 2005–2007, Warszawa 2008.

²⁰ Por. J. Janczyk, *Racjonalność użytkowania Internetu* [w:] *Racjonalność myślenia, decydowania i działania*, red. L.W. Zacher, Wyd. WSPiZ, Warszawa 2000.

- Ferguson R., *The State Of Learning Analytics in 2012: A Review and Future Challenges*, Technical Report KMI-12-01, Knowledge Media Institute, The Open University, UK 2012, <http://kmi.open.ac.uk/publications/techreport/kmi-12-01> (dostęp: 21.01.2013 r.).
- Janczyk J., *Netykieta w środowisku nauczycieli przedmiotów informatycznych* [w:] *Historia, instytucjonalizacja i perspektywy kształcenia nauczycieli na Śląsku – Jubileusz 80-lecia*, red. S. Juszczyk, D. Morańska, Wyd. Uniwersytet Śląski, Katowice 2010.
- Janczyk J., *Racjonalność użytkowania Internetu* [w:] *Racjonalność myślenia, decydowania i działania*, red. L.W. Zacher, Wyd. WSPiZ, Warszawa 2000.
- Janczyk J., *Technical and Organisational Crises in Nets*, TRANSFORMACJE Special Issue 2005–2007, Warszawa 2008.
- Kreft P., *Większość ruchu w internecie generują... boty*, Komputer Świat, <http://www.komputer-swiat.pl/nawosci/internet/2013/50/wiekszosc-ruchu-w-internecie-generuja-boty.aspx> (dostęp: 14.12.2013 r.).
- Manyika J., Chui M., Brown B., Bughin J., Dobbs R., Roxburgh C., Hung Byers A., *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, McKinsey report: McKinsey Global Institute 2011.
- Mayer-Schonberger V., Cukier K., *Big Data Rewolucja która zmieni nasze myślenie, pracę i życie*, Wyd. MT Biznes, Warszawa 2014.
- Romero C., Ventura S., *Educational Data Mining: A Review of the State of the Art*, IEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews 40(6)/2010.
- Zikopoulos P.C., Eaton C., deRoos D., Deutsch T., Lapis G., *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*, McGraw-Hill books, USA 2012.