Jacek Maślankowski

Wydział Zarządzania
Uniwersytet Gdański

Analiza jakości danych pozyskiwanych ze stron internetowych z wykorzystaniem rozwiązań *Big Data*

1. Wstęp

Podczas przygotowywania badania statystycznego dużą wagę przywiązuje się do jakości danych. W wynikach badań reprezentacyjnych są sprawdzane w szczególności błędy standardowe. W przypadku odchyleń dla błędów względnych, najczęściej przekraczających 10%, uznaje się, że dane są mało wiarygodne.

Coraz częściej w literaturze można zetknąć się z zastosowaniem systemów *Big Data* do wspierania realizacji różnego rodzaju badań, w tym statystycznych. Celem jest przede wszystkim zmniejszenie kosztów związanych z realizacją badania tradycyjną metodą, tj. przez sieć ankieterską lub zlecając realizację podmiotom zewnętrznym.

Pytanie zatem brzmi: czy w przypadku rozwiązań *Big Data* można mówić o danych dobrej jakości? Celem niniejszego artykułu jest zbadanie możliwości wyprodukowania dobrej jakości danych statystycznych pozyskanych dzięki zastosowaniu tych systemów. W pierwszej części przedstawiono przesłanki związane z wykorzystaniem systemów *Big Data* do badania różnego rodzaju zjawisk. Druga część obrazuje istotę zastosowań tych systemów i możliwość ich szerokiego zastosowania również przez administrację publiczną. W trzeciej części zawarto definicję jakości danych z wykorzystaniem systemów *Big Data*. W czwartej części znajduje się opis rozwiązania testowego oraz wyniki przeprowadzonego pilotażu oceny jakości danych pozyskiwanych z tych systemów. Piąta część prezentuje propozycję szablonu jakości danych w kontekście przetwarzania ich z zastosowaniem rozwiązań *Big Data*. W ostatniej części zostały zamieszczone wnioski i plany dalszych prac w tym obszarze.

2. *Big Data* w badaniach naukowych

Analizując artykuły naukowe z minionych 5 lat, można zaobserwować znaczący wzrost zainteresowania w nauce zastosowaniem rozwiązań *Big Data*, których głównym celem jest umożliwianie równoległego przetwarzanie dużych zbiorów danych. Dotyczy to wielu różnego rodzaju zastosowań, począwszy od przetwarzania dużych zbiorów danych ustrukturyzowanych, na pozyskiwaniu wiedzy z informacji nieustrukturyzowanej skończywszy. O znaczeniu tego rodzaju rozwiązań może świadczyć liczba artykułów, które ukazały się w znaczących opracowaniach naukowych. Zostałay one zaprezentowane w tabeli 1.

Tabela 1. Statystyka recenzowanych artykułów w czasopismach naukowych

Wyszczególnienie	Liczba artykułów w roku				
	2009	2010	2011	2012	2013
Baza danych Business Sources Complete (tylko recenzowane artykuły naukowe)	2	1	6	47	251
Baza danych Business Sources Complete (łączna liczba artykułów)	7	12	114	551	1250

Źródło: opracowanie własne.

Uzupełniając informacje zawarte w tabeli, należy nadmienić, że łączna liczba artykułów związanych z tematyką *Big Data* w bazie Business Source Complete przekroczyła obecnie 3 tys.

Big Data można zatem określać w kategoriach instrumentu naukowego nowej generacji. Mogą to być pozornie proste analizy, takie jak badanie zachowań konsumentów na podstawie informacji zamieszczanych w Internecie, jak również złożone algorytmy w zakresie zaawansowanych obliczeń np. rozproszonych fotonów[1]. W powszechnej opinii rozwiązania tego typu są jednak utożsamiane z możliwością przetwarzania informacji nieustrukturyzowanej, pochodzącej ze stron WWW, np. z serwisów społecznościowych, takich jak Twitter, Facebook i innych[2]. Należy zaznaczyć, że ponad 90% informacji na świecie jest zapisywane w formie nieustrukturyzowanej[3].

---

[1] A. Wright, *Big Data Meets Big Science*, „Communications of the ACM" 2014, vol. 57, issue 7, July, s. 13–15.

[2] T.K. Das, P.M. Kumar, *BIG Data Analytics: A Framework for Unstructured Data Analysis*, „International Journal of Engineering Science & Technology" 2013, vol. 5, issue 2, February, s. 153–156; P. Płoszajski, *Big Data: nowe źródło przewag i wzrostu firm*, „E-mentor" 2013, nr 3(50), s. 5–10.

[3] K. Gang-Hoon, S. Trimi, C. Ji-Hyong, *Big-Data Applications in the Government Sector*, „Communications Of The ACM" 2014, vol. 57, no. 3, s. 78–85.

Analizując literaturę w tym zakresie, zarówno badacze ze świata nauki, jak i praktycy podkreślają, że *Big Data* pozwala analizować chociażby zachowania konsumentów na bardzo dużej próbie badawczej. Jeżeli zatem dla Polski przyjmuje się próbę wielkości 1000 osób, to wydawać się może, że analizowanie treści pisanych przez miliony Polaków na Twitterze pozwala lepiej zrozumieć zachowanie, nastawienie czy preferencje społeczeństwa względem określonych produktów czy zjawisk społecznych.

Big Data nie jest nową technologią, a jedynie umożliwia wykorzystywanie powszechnie znanych metod statystyki opisowej czy wnioskowania statystycznego dla zbiorów ustrukturyzowanych o bardzo dużych rozmiarach. W bardziej zaawansowanej formie pozwala na zastosowanie metod *text mining* dla dużych zbiorów nieustrukturyzowanych[4] lub integrację źródeł informacji nieustrukturyzowanej i ustrukturyzowanej[5].

Zatem należy zadać pytanie: czy *Big Data* może zastąpić badanie prowadzone tradycyjną metodą? Hipoteza postawiona w niniejszym artykule jest następująca: rozwiązania *Big Data* pozwalają na uzyskanie danych dobrej jakości pod warunkiem prawidłowego przygotowania algorytmu analizy składni stron internetowych, które mają zasilać bazę danych wynikowych.

3. Jakość danych według ich przeznaczenia

Jakość danych jest przedmiotem rozważań środowisk naukowych oraz licznych instytucji, które wydają stosowne dokumenty normatywne, w tym zarządzenia mające na celu poprawę jakości przetwarzanych danych. Jakość często jest definiowana zgodnie z normą ISO 8402:1986 z późniejszymi zmianami. Według tej normy, jakość to „ogół cech i właściwości produktu lub usługi, który decyduje o zdolności zaspokojenia potrzeb zadeklarowanych lub domyślnych”[6]. W normie ISO-9001:2008 system zarządzania jakością został zdefiniowany w ośmiu zasadach jakości, uwzględniających m.in. zorientowanie na klienta oraz podejścia procesowe i systemowe do zarządzania.

Często pojęcie jakości danych jest utożsamiane z ich dokładnością. Jednak jest to zbyt uproszczony sposób traktowania danych, gdyż pojęcie to warunkują również dwa dodatkowe kryteria: przydatność danych, oznaczająca ich relewantny charakter

[4] E.W. Kuiler, *From Big Data to Knowledge: An Ontological Approach to Big Data Analytics*, „Review of Policy Research” 2014, vol. 31, issue 4, July, s. 311–318.

[5] J. Maślankowski, *Data Quality Issues Concerning Statistical Data Gathering Supported by Big Data Technology*, „Communications in Computer and Information Science” 2014, no. 424, s. 92–101.

[6] http://www.stat.gov.pl/gus/5466_PLK_HTML.htm.

dla odbiorcy, oraz aktualność danych, czyli to, że powinny być dostarczane w określonym czasie[7].

W polskiej statystyce publicznej jakość danych jest definiowana zgodnie z europejskim systemem statystycznym przez rozpatrywanie sześciu komponentów jakości: przydatności, dokładności, terminowości i punktualności, dostępności i przejrzystości, porównywalności, spójności[8].

Z kolei w analityce biznesowej jakość danych jest utożsamiana z kompletnością i spójnością, jako centralnymi wymiarami jakości danych systemów analitycznych[9]. Rozpatrywanie jakości danych biznesowych powinno odbywać się w trzech etapach – na wejściu, podczas przetwarzania oraz na wyjściu[10].

Bardziej złożoną charakterystykę jakości danych można znaleźć w opracowaniach naukowych. Jako przykład może posłużyć badanie na temat jakości danych przeprowadzone wśród 25 pracowników oraz 112 studentów MBA w Stanach Zjednoczonych. Celem badania było zdefiniowanie przez respondentów atrybutów jakości danych. W taki sposób uzyskano 179 atrybutów, które są postrzegane przez badane osoby jako kryteria oceny jakości danych zwane wymiarami. Wśród nich jako najważniejsze sklasyfikowano: wiarygodność danych, wartość dodaną, relewantność, dokładność oraz interpretowalność. Wymiary te zostały przyporządkowane do czterech kategorii: dokładność, relewantność, reprezentatywność (w znaczeniu interpretowalności, spójności oraz łatwości zrozumienia) oraz dostępność[11]. Inna perspektywa jakości danych uwzględnia cztery główne wymiary: kompletność, jednoznaczność, znaczenie oraz poprawność[12].

Abstrahując od powyższych istotnych kryteriów oceny jakości danych, należy stwierdzić, że ważnym czynnikiem mającym wpływ na wysoką jakość danych statystycznych jest dobór próby do badania. Dlatego przy badaniach statystycznych należy zwykle znaleźć kompromis pomiędzy jakością danych statystycznych a kosztem

---

7 J. Kordos, *Dokładność danych w badaniach społecznych*, „Biblioteka Wiadomości Statystycznych" (GUS) 1987, t. 35.

8 Ibidem.

9 O. Kwon, N. Lee, B. Shin, *Data quality management, data usage experience and acquisition intention of big data analytics*, „International Journal of Information Management" 2014, vol. 34, issue 3, June, s. 387–394.

10 B. Hazen, C. Boone, J. Ezell, L.A. Jones-Farmer, *Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications*, „International Journal of Production Economics" 2014, vol. 154, August, s. 72–80.

11 R.Y. Wang, D.M. Strong, *Beyond Accuracy: What Data Quality Means to Data Consumers,* „Journal of Management Information Systems" 1996, vol. 12, no. 4, s. 5–34.

12 A. Haug, J.S. Arlbjorn, F. Zachariassen, J. Schlichter, *Master data quality barriers: an empirical investigation*, „Industrial Management & Data Systems" 2014, vol. 113, no. 2, s. 234–249.

przeprowadzenia badania. Do tego celu wykorzystuje się metodę reprezentacyjną[13]. Podsumowując: zgodnie z zasadą GIGO (ang. *garbage-in garbage-out*), zgromadzenie danych złej jakości przyczyni się do uzyskania błędnych wyników.

Niektóre środowiska traktują jakość danych w sposób wielowymiarowy. Przykładem jest chociażby próba zdefiniowania jakości danych dla rozwiązań *Big Data* przez Europejską Komisję Gospodarczą ONZ[14]. W ramach tej metody wyróżnia się trzy podstawowe duże wymiary jakości danych (ang. *hyperdimensions*):

- dane;
- metadane;
- źródła.

Te trzy duże wymiary są rozpatrywane na trzech poziomach: wejście, przetwarzanie, wyjście. Poziom wejściowy jest związany z pozyskiwaniem danych. Poziom przetwarzania danych odzwierciedla procesy zachodzące po pozyskaniu danych, jeszcze przed opublikowaniem wyników. Poziom wyjściowy służy do określenia jakości danych wynikowych. I tak, dla oceny źródeł danych następuje weryfikacja aspektów prywatności i zabezpieczeń oraz otoczenia biznesowego. Dla metadanych najczęściej jest oceniana złożoność, kompletność oraz użyteczność. Dla danych jest badana dokładność i możliwość jednoznacznego łączenia danych w grupy.

Oceniając jakość danych *Big Data*, należy również zwrócić uwagę na cykl życia danych, który obejmuje etapy od generowania danych, przez pozyskiwanie, gromadzenie do przetwarzania (analizy)[15]. Model pomiaru jakości oprogramowania *Big Data* powinien składać się z co najmniej czterech etapów: definiowania wymagań jakości, pomiaru, agregacji i analizy wyników[16].

Nadmierna liczba atrybutów dotyczących jakości danych może prowadzić do problemów z jednoznacznym określeniem wartości tych atrybutów. Szczególnie dotyczy to atrybutów, które nie mają dokładnie zdefiniowanych mierników.

4. Środowisko testowe i wyniki analiz

Aby potwierdzić tezę badawczą zawartą w niniejszym artykule, zastosowano środowisko testowe składające się z kilku komponentów. Są to przede wszystkim: robot

[13] M. Szreder, *Metody i techniki sondażowych badań opinii*, PWE, Warszawa 2004.

[14] Informacje pochodzą z konferencji na temat *Big Data* w badaniach statystycznych.

[15] M. Chen, S. Mao, Y. Liu, *Big Data: A Survey*, „Mobile Network Applications" 2014, no. 19, s. 171–209.

[16] G. Suciu, A. Vulpe, G. Todoran, T.-L. Militaru, *Cloud Computing and Big Data as Convergent Technologies for Mobile E-Learning*, „Elearning & Software for Education" 2014, issue 1, s. 113–120.

zbierający dane, który jest oprogramowaniem pozwalającym na pobieranie danych, oraz oprogramowanie do równoległego przetwarzania danych. Proponowane rozwiązanie wykorzystuje oprogramowanie Apache Hadoop, będące jednym z najczęściej polecanych systemów do równoległego przetwarzania dużych zbiorów danych[17]. Wraz z modelem programowania MapReduce pozwala na znajdowanie wzorców na zasadzie klucz-wartość (ang. *key-value*)[18]. Wyniki analiz są gromadzone w specjalnej strukturze hurtowni danych[19]. Ponieważ niniejszy artykuł nie skupia się na zagadnieniach technicznych, dlatego szerzej o sposobie przechowywania danych w taki sposób można przeczytać w innych artykułach autora niniejszego opracowania.

Analizy określające możliwości oceny jakości danych *Big Data* przeprowadzono, bazując na portalach internetowych zawierających oferty pracy. Celem było zbadanie popytu na pracę przez sprawdzenie, jakie zawody są obecnie najbardziej pożądane na rynku pracy. Procedura testowania danych odbyła się w następujący sposób: 1) pobranie danych ze strony internetowej; 2) przetwarzanie danych; 3) analiza otrzymanych wyników. Zdecydowano się na przeprowadzenie trzech testów związanych z indeksowaniem i przetwarzaniem wyników standardowymi narzędziami *Big Data*. Założenia do tych testów zostały opisane w tabeli 2.

Tabela 2. Rodzaje przeprowadzonych testów dotyczących przetwarzania danych *Big Data*

Numer testu	Opis	Charakterystyka
1	analiza wszystkich słów kluczowych	duże raporty analityczne, przypadkowość wyników
2	analiza ściśle określonych słów kluczowych	analityka dobrana do potrzeb, możliwość pominięcia istotnych informacji
3	analiza powiązań pomiędzy słowami kluczowymi i innymi atrybutami	szerszy obraz analiz, możliwość znalezienia nieznanych dotychczas powiązań

Źródło: opracowanie własne.

W pierwszym teście pobrano strony internetowe z jednego z najpopularniejszych polskich portali internetowych zamieszczających oferty pracy. Aby zobrazować skalę zjawiska związanego z otrzymywaniem nieprawidłowych danych, pobrano niewielką liczbę danych, tj. 172 kB danych tekstowych, do przeanalizowania.

[17] G. Mone, *Beyond Hadoop*, „Communications of the ACM" 2013, vol. 56, issue 1, January, s. 22–24.

[18] C. Boja, A. Pocovnicu, L. Bătăgan, *Distributed Parallel Architecture for "Big Data"*, „Informatica Economica" 2012, vol. 16, issue 2, s. 116–127.

[19] J. Maślankowski, *The integration of web-based information and the structured data in data warehousing*, „Lecture Notes in Business Information Processing" 2013, no. 161, s. 66–75.

Wykorzystany w pierwszym teście algorytm MapReduce zwrócił wynik zawierający 1332 wiersze analiz słów kluczowych. Analizując wyniki i porównując je ze stanem faktycznym, tj. popytem na pracę, należy uznać, że blisko 10% wierszy jest oczekiwanym rezultatem. Pozostałe 90% należałoby odrzucić, ze względu na przypadkowość dokonanych analiz.

W drugim teście algorytm uwzględniał jedynie te słowa kluczowe, które zostały podane w wymaganiach, np. zawód programista. Głównym problemem w tym teście była możliwość pominięcia istotnych informacji, które mogłyby bardziej charakteryzować dany zawód. Dodatkowo istotnym problemem okazywały się duplikaty danych.

Trzeci test dotyczył analiz słów kluczowych w sposób krzyżowy, tj. znajdowania powiązań pomiędzy słowami kluczowymi i zdefiniowanymi wcześniej atrybutami będącymi metadanymi na stronie internetowej. Test ten jednak nie sprawdził się w praktyce, gdyż liczba wykrywanych zależności znacząco przekraczała możliwość automatycznego wygenerowania syntetycznego ujęcia.

5. Proponowane atrybuty oraz mierniki jakości danych *Big Data*

Jak wspomniano we wcześniejszej części artykułu, zagadnienia jakości danych powinny zostać podzielone na trzy grupy: wejściowe, przetwarzania oraz wyjściowe. Te pierwsze dotyczą oceny zbioru źródłowego danych przed jego pobraniem. Drugie mają na celu ocenę jakości i analizują wyniki przetwarzania tego zbioru danych. Trzecie dotyczą oceny zbioru danych wyjściowych.

Zastosowana w niniejszym artykule metoda badawcza polegała na przeprowadzeniu eksperymentu związanego z wykorzystaniem danych nieustrukturyzowanych jako źródeł *Big Data*. Przetwarzanie danych potwierdziło powszechnie znaną tezę, że o jakości danych wynikowych w głównej mierze decyduje źródło danych i sposób ich przetwarzania.

Przeprowadzone testy doprowadziły do wniosków, że kluczowe są następujące wymiary jakości danych *Big Data*:

- jednoznaczność;
- obiektywizm (błędy tzw. mapowania i redukowania);
- zawieranie znacznika czasowego;
- granularność;
- występowanie duplikatów danych;
- kompletność;
- dostępność;

- precyzja;
- interpretowalność;
- integralność;
- spójność.

W powyższym zestawieniu celowo pominięto ogólne wymiary jakości danych, takie jak dokładność, reprezentacyjność czy aktualność. Skoncentrowano się jedynie na tych wymiarach, które są specyficzne w procesie przetwarzania danych *Big Data*.

Bazując na przeprowadzonej analizie danych oraz wynikach testów opisanych we wcześniejszych punktach niniejszego opracowania, można zdefiniować mierniki określające jakość danych *Big Data* (tabela 3).

Tabela 3. Proponowane kluczowe mierniki jakości źródeł danych *Big Data*

Lp.	Nazwa miernika i sposób reprezentacji	Miara
1.	Struktura – czy dane źródłowe są ustrukturyzowane, nieustrukturyzowane czy też częściowo ustrukturyzowane?	procent kategorii
2.	Format danych – czy jest to standardowy format danych, np. język znaczników, plik binarny, dokument tekstowy?	procent kategorii
3.	Wykorzystane standardy opisu danych – jak wiele klasyfikacji wykorzystanych w zbiorze danych można zidentyfikować, np. ISO-3166 dla opisu krajów, PKD dla opisu działalności, ISCED dla opisu kierunków studiów czy KZiS dla opisu zawodów?	lista standardów
4.	Złożoność – jak wiele plików lub tabel tworzy źródło danych i czy są one zunifikowane?	liczba bezwzględna z podziałem na kategorie

Źródło: opracowanie własne.

Zawarte w tabeli 3 mierniki obrazują kluczowe jedynie w kontekście *Big Data* mierniki. Należy pamiętać o tym, aby analizując przydatność źródła danych, uwzględniać również takie mierniki, jak dokładność, kompletność, dostępność i inne powszechnie przyjęte do oceny jakości źródeł danych.

W tabeli 4 zawarto mierniki jakości danych *Big Data*, które podczas przetwarzania danych mogą zostać wykorzystane do opisu jakości przetwarzania danych. Mierniki te mają jednoznaczną interpretację. W zależności od miernika, im wyższa lub niższa jest wartość, tym źródło danych mniej lub bardziej nadaje się do zastosowania w procesie przygotowywania danych wynikowych.

Proponowane mierniki do oceny jakości danych mogą zostać zastosowane na etapie przetwarzania danych. Stanowi to wadę, gdyż przygotowanie algorytmów MapReduce, zastosowanych w testach nr 2 i 3, jest czasochłonne i w przypadku wielu źródeł danych nakład czasu pracy programisty może być znaczący.

Tabela 4. Proponowane kluczowe mierniki jakości danych *Big Data*

Lp.	Nazwa miernika i sposób reprezentacji	Miara
1.	Jednoznaczność – jak wiele obserwacji (rekordów) nie jest jednoznacznych i może być interpretowanych w wieloraki sposób?	procent
2.	Obiektywizm – jak dużo rekordów może być zamapowanych niewłaściwie?	procent
3.	Zawieranie znacznika czasowego – jak wiele obserwacji można przyporządkować do określonego znacznika czasowego?	procent
4.	Granularność lub stopień szczegółowości – jak wiele obserwacji może być przyporządkowanych do wielu poziomów hierarchii, np. skala wojewódzka, powiatowa itd.?	procent
5.	Występowanie duplikatów danych – ile duplikatów zostało zidentyfikowanych w procesie przetwarzania danych?	procent
6.	Kompletność – jaki odsetek obserwacji, rekordów typu klucz-wartość zostało odrzuconych w procesie przetwarzania?	procent
7.	Dostępność – jaki odsetek obserwacji został pozyskany ze źródła bez żadnych błędów i ostrzeżeń wynikających z działania algorytmu podczas procesu przetwarzania?	procent
8.	Zgodność ze wzorcem – jak wiele obserwacji nie może być zamapowanych jako klucz-wartość zgodnie z przyjętym wzorcem czy wyrażeniem regularnym?	procent
9.	Klasyfikacja i kategoryzacja – jak wiele obserwacji nie może zostać przypisanych do przyjętych klasyfikacji i kategorii?	procent
10.	Precyzja – jak wiele obserwacji podczas procesu dopasowywania do wzorca traci część swojego opisu?	procent
11.	Interpretowalność – ile rekordów może być błędnie zinterpretowanych?	procent
12.	Integralność – jak wiele obserwacji jest zmienianych w procesie przetwarzania?	procent
13.	Spójność – jak wiele obserwacji zostało przekształconych lub przekonwertowanych podczas procesu przetwarzania danych?	procent

Źródło: opracowanie własne.

6. Podsumowanie i kierunki dalszych badań

Przeprowadzona analiza potwierdza przyjętą wcześniej hipotezę, że istnieje możliwość uzyskania wyników dobrej jakości na podstawie informacji nieustrukturyzowanej, przetwarzanej z zastosowaniem rozwiązań *Big Data*. Należy jednak dobierać źródła danych w sposób umożliwiający pomiar jakości danych. Ważne jest uzyskanie satysfakcjonujących wartości mierników zaprezentowanych w niniejszym opracowaniu w tabelach 3 i 4. Poza wymienionymi miernikami, istotnym zagadnieniem związanym z wyborem źródła danych stanowi jego reprezentacyjność i kompletność. Tylko w taki sposób istnieje możliwość uzyskania wiarygodnych wyników. Wpływ na wiarygodność danych będzie miał w tym przypadku przede wszystkim operat badania.

Kierunki przyszłych badań będą związane przede wszystkim z dalszym rozwojem modelu badania jakości danych *Big Data* i jego weryfikacją w ramach pracy z danymi ustrukturyzowanymi.

Bibliografia

Boja C., Pocovnicu A., Bătăgan L., *Distributed Parallel Architecture for "Big Data"*, „Informatica Economica" 2012, vol. 16, issue 2, s. 116–127.

Chen M., Mao S., Liu Y., *Big Data: A Survey*, „Mobile Network Applications" 2014, no. 19, s. 171–209.

Das T.K., Kumar P.M., *BIG Data Analytics: A Framework for Unstructured Data Analysis*, „International Journal of Engineering Science & Technology" 2013, vol. 5, issue 2, February, s. 153–156.

Gang-Hoon K., Trimi S., Ji-Hyong C., *Big-Data Applications in the Government Sector*, „Communications Of The ACM" 2014, vol. 57, no. 3, s. 78–85.

Haug A., Arlbjorn J.S., Zachariassen F., Schlichter J., *Master data quality barriers: an empirical investigation*, „Industrial Management & Data Systems" 2014, vol. 113, no. 2, s. 234–249.

Hazen B., Boone C., Ezell J., Jones-Farmer L.A., *Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications*, „International Journal of Production Economics" 2014, vol. 154, August, s. 72–80.

Kordos J., *Dokładność danych w badaniach społecznych*, „Biblioteka Wiadomości Statystycznych" (GUS) 1987, t. 35.

Kuiler E.W., *From Big Data to Knowledge: An Ontological Approach to Big Data Analytics*, „Review of Policy Research" 2014, vol. 31, issue 4, July, s. 311–318.

Kwon O., Lee N., Shin B., *Data quality management, data usage experience and acquisition intention of big data analytics*, „International Journal of Information Management" 2014, vol. 34, issue 3, June, s. 387–394.

Maślankowski J., *Data Quality Issues Concerning Statistical Data Gathering Supported by Big Data Technology*, „Communications in Computer and Information Science" 2014, no. 424, s. 92–101.

Maślankowski J., *The integration of web-based information and the structured data in data warehousing*, „Lecture Notes in Business Information Processing" 2013, no. 161, s. 66–75.

Mone G., *Beyond Hadoop*, „Communications of the ACM" 2013, vol. 56, issue 1, January, s. 22–24.

Płoszajski P., *Big Data: nowe źródło przewag i wzrostu firm*, „E-mentor" 2013, nr 3(50), s. 5–10.

Suciu G., Vulpe A., Todoran G., Militaru T.-L., *Cloud Computing and Big Data as Convergent Technologies for Mobile E-Learning*, „Elearning & Software for Education" 2014, issue 1, s. 113–120.

Szreder M., *Metody i techniki sondażowych badań opinii*, PWE, Warszawa 2004.

Wang R.Y., Strong D.M., *Beyond Accuracy: What Data Quality Means to Data Consumers*, „Journal of Management Information Systems" 1996, vol. 12, no. 4, s. 5–34.

Wright A., *Big Data Meets Big Science*, „Communications of the ACM" 2014, vol. 57, issue 7, July, s. 13–15.

Źródła sieciowe

http://www.stat.gov.pl/gus/5466_PLK_HTML.htm (data odczytu: 24.08.2014).

* * *

Big Data quality analysis on data retrieved from websites

Summary

The article presents a proposition of a Big Data quality framework in terms of processing Big Data sources to produce statistical information. The case used in the article concerns job offers that generate information about the demand of the labour market. The analyses has resulted in a suggestion of several quality dimensions with indicators.

Keywords: Big Data, data quality, unstructured data