

**Wykorzystanie
metod**

BIG DATA

**do badania organizacji
gospodarczych obecnych
w Internecie**

Jacek Maślankowski

Wykorzystanie
metod

BIG DATA



do badania organizacji
gospodarczych obecnych
w Internecie

Teoria, metodologia i wyniki

Wydawnictwo Uniwersytetu Gdańskiego
Gdańsk 2025

Recenzenci
dr hab. Maria Mach-Król
dr hab. Jakub Swacha

Redaktor Wydawnictwa
Magdalena Cichowicz-Cieślak

Projekt okładki i stron tytułowych
Angram Studio Weronika Jastrzemska

Skład i łamanie
Mariusz Szewczyk

Publikacja dofinansowana ze środków
Prorektora Uniwersytetu Gdańskiego ds. Badań Naukowych
w ramach Programu Humanistyki Gdańskiej

© Copyright by Uniwersytet Gdański
Wydawnictwo Uniwersytetu Gdańskiego

ISBN 978-83-8206-744-6

Wydawnictwo Uniwersytetu Gdańskiego
ul. Armii Krajowej 119/121, 81-824 Sopot
tel. +48 58 523 11 37, tel. kom. +48 725 991 206
e-mail: wydawnictwo@ug.edu.pl
wydawnictwo.ug.edu.pl

Księgarnia internetowa: wydawnictwo.ug.edu.pl/sklep/

Druk i oprawa
Zakład Poligrafii Uniwersytetu Gdańskiego
ul. Armii Krajowej 119/121, 81-824 Sopot
tel. +48 58 523 14 49

Spis treści

Stosowane skróty i akronimy	9
Wprowadzenie	13
Rozdział 1. Pojęcie i zastosowania <i>big data</i>	17
1.1. Przegląd literatury naukowej	17
1.1.1. Koncepcja i rozwój <i>big data</i>	17
1.1.2. Podstawowe definicje	21
1.1.3. Zastosowanie <i>big data</i> w praktyce	22
1.2. Rozwój systemów informatycznych wspierających przetwarzanie dużych zbiorów danych	24
1.3. Krytyka podejść związanych ze stosowaniem <i>big data</i>	26
1.4. Strony internetowe jako źródło danych biznesowych	26
Rozdział 2. Typologia danych <i>big data</i>	29
2.1. Podział danych	29
2.2. Rodzaje danych ze względu na źródło powstawania	30
2.3. Stopień ustrukturyzowania danych	34
Rozdział 3. Metodologiczne aspekty badań <i>big data</i>	37
3.1. Wybrane metody związane z analizą dużych zbiorów danych	37
3.2. Metody wdrożeń rozwiązań wykorzystujących źródła <i>big data</i>	40
3.3. Model biznesowy sterowany danymi	42
3.4. Próba badawcza	44
3.4.1. Ogólne uwagi dotyczące próbkowania danych	44
3.4.2. Próby nielosowe	45
3.5. Architektura systemowa i procesowa	45
3.6. Metody statystyczne	49
3.7. Metody uczenia maszynowego	51
3.8. Metody <i>web mining</i>	56
3.8.1. Rodzaje i zastosowanie	56
3.8.2. Model Bag-of-words	56
3.8.3. Przykłady Vector Space: Document Term Matrix oraz Term Document Matrix	58

3.8.4. Metoda TF-IDF	59
3.9. Łączenie zbiorów danych	60
3.10. Zarządzanie zmianami w źródłach danych	62
3.11. Badania tradycyjne a <i>big data</i>	63
Rozdział 4. Zagadnienia jakości danych typu <i>big data</i>	65
4.1. Przegląd podejść stosowanych do oceny jakości danych	65
4.2. Wymiary jakości danych	69
4.3. Syntetyczne ujęcie mierników jakości danych	76
4.4. Przykład badania jakości danych w przypadku klasyfikowania z wykorzystaniem uczenia maszynowego	78
4.5. Sugerowane podejście do oceny jakości danych typu <i>big data</i>	80
Rozdział 5. <i>Big data</i> w badaniach naukowych bazujących na danych z Internetu	83
5.1. <i>Big data</i> a dziedziny nauki	83
5.2. <i>Big data</i> w środowisku naukowym	87
5.2.1. Ośrodki naukowe <i>big data</i>	87
5.2.2. Przykłady zastosowań <i>big data</i> według dziedzin nauki	88
5.3. Przegląd badań korzystających z metod i źródeł <i>big data</i>	89
5.3.1. Badania edukacyjne	89
5.3.2. Badania rynku pracy	92
5.3.3. Badania serwisów społecznościowych	95
5.3.4. Badania językowe	98
5.3.5. Badania trendów	98
5.3.6. Badania społeczne	102
5.3.7. Badania finansowe	103
5.3.8. Badania środowiska	105
5.3.9. Badania przedsiębiorstw	106
5.4. Wyzwania w obszarze nauki związane z <i>big data</i>	106
Rozdział 6. Badanie stron internetowych organizacji działających w Polsce	109
6.1. Metodyka badania	109
6.1.1. Wprowadzenie	109
6.1.2. Wyszukiwanie adresów stron internetowych	110
6.1.3. Pobieranie zawartości stron internetowych	111
6.1.4. Identyfikacja kont w mediach społecznościowych	113
6.1.5. Pobieranie danych dotyczących ofert pracy	114
6.1.6. Przetwarzanie i analiza danych ze stron internetowych	115
6.1.7. Sugerowana metodyka prowadzenia badań na podstawie stron internetowych	116
6.2. Opis badania	120
6.2.1. Próba badania	120
6.2.2. Hipotezy badawcze	123
6.2.3. Charakterystyka czasowa i przestrzenna	124

6.3. Walidacja danych	125
6.4. Ocena jakości danych	125
Rozdział 7. Rezultaty badania dotyczące charakterystyki organizacji gospodarczych w Polsce	
7.1. Podział wniosków badania	127
7.2. Zainteresowanie terminem <i>big data</i> na badanym obszarze	128
7.3. Wielkość jednostek obecnych w Internecie oraz rodzaj ich działalności	129
7.4. Atrybuty stron internetowych firm	137
7.4.1. Obecność pliku robots.txt	137
7.4.2. Język stron internetowych	138
7.4.3. Analiza treści strony na podstawie słów kluczowych	140
7.4.4. Analiza łączy do mediów społecznościowych	142
7.4.5. Analiza ofert pracy związanych z <i>big data</i>	151
7.4.6. Podział stron internetowych na klastry z wykorzystaniem uczenia nienadzorowanego	153
7.4.7. Rozpoznawanie innowacyjnej działalności z wykorzystaniem uczenia maszynowego nadzorowanego	156
7.4.8. Odniesienie do technologii <i>big data</i> na stronach internetowych	159
7.5. Metody i wyniki walidacji danych	161
7.5.1. Adresy URL	161
7.5.2. Obecność pliku robots.txt	162
7.5.3. Język stron internetowych	163
7.5.4. Analiza treści strony na podstawie słów kluczowych	163
7.5.5. Analiza łączy do mediów społecznościowych	164
7.5.6. Analiza ofert pracy związanych z <i>big data</i>	165
7.5.7. Podział stron internetowych na klastry z wykorzystaniem uczenia nienadzorowanego	166
7.5.8. Rozpoznawanie innowacyjnej działalności z wykorzystaniem uczenia maszynowego nadzorowanego	166
7.5.9. Odniesienie do technologii <i>big data</i> na stronach internetowych	167
7.6. Dyskusja dotycząca wyzwań związanych z prowadzeniem badań na podstawie źródeł <i>big data</i> w postaci stron internetowych	168
7.6.1. Próba badawcza	168
7.6.2. Prawne aspekty pozyskiwania danych z Internetu	170
7.7. Podsumowanie wyników badań	171
Zakończenie	175
Bibliografia	177
Spis rysunków	195
Spis tabel	197

Stosowane skróty i akronimy

3V – Volume, Veracity, Velocity
5V – Volume, Velocity, Variety, Veracity, Value
AI – Artificial Intelligence
API – Application Programming Interface
B2C – Business-To-Customer
BAEL – Badanie aktywności ekonomicznej ludności
BDAQ – Big Data Analytics Quality
BERT – Bidirectional Encoder Representations from Transformers
BI – Business Intelligence
BI&A – Business Intelligence & Analytics
BN – Bayesian Network (uczenie maszynowe)
BPR – Bayesian Profile Regression
CART – Classification and Regression Trees (uczenie maszynowe)
CAWI – Computer Assisted Web Interviewing
CB – Competitive Benchmarking
CCI – Consumer Confidence Index
CEDEFOP – Centre Européen pour le Développement de la Formation Professionnelle lub
European Centre for the Development of Vocational Training – Europejskie Centrum
Rozwoju Kształcenia Zawodowego
CRM – Customer Relationship Management
DA – Discriminant Analysis (uczenie maszynowe)
DJIA – Dow Jones Industrial Average
DL – Deep Learning
DQ – Data Quality
DT – Decision Trees (uczenie maszynowe)
EDM – Educational Data Mining
EKG ONZ –Europejska Komisja Gospodarcza Organizacji Narodów Zjednoczonych (patrz
UNECE)
EMIS – Emerging Markets Information Service
EPA – The US Environmental Protection Agency – Agencja Ochrony Środowiska Stanów
Zjednoczonych
ERP – Enterprise Resource Planning
ESCO – European Standard Classification of Occupation
ESS – European Social System

ESSnet – European Social System network (sieć krajów Europejskiego Systemu Statystycznego w zakresie współpracy międzynarodowej)

FTSE – Financial Times Stock Exchange

GFS – Google File System

GP – Gaussian Process (uczenie maszynowe)

GSBPM – Generic Statistical Business Process Model

HDFS – Hadoop Distributed File System

HHI – Herfindahl–Hirschman Index lub indeks Herfindahla–Hirschmana

ICT – Information and Communication Technology

IF – Impact Factor

IoT – Internet of Things

ISCED – International Standard Classification of Education

ISCO – International Standard Classification of Occupation

ISO – International Organization for Standardization

IT – Information Technology

KZiS – Klasyfikacja Zawodów i Specjalności

LMC – Labour Market Concentration

LR – Logistic Regression (uczenie maszynowe)

MIS – Management Information Systems

ML – Machine Learning

MTBF – Mean Time Between Failures

MTTF – Mean Time to Failures

MTTR – Mean Time to Repair

N225 – Nikkei 225

NACE – Nomenclature statistique des activités économiques dans la Communauté européenne lub Statistical Classification of Economic Activities in the European Community

NB – Naïve Bayes (uczenie maszynowe)

NIP – Numer Identyfikacji Podatkowej

NLP – Natural Language Processing

NN – Neural Network (uczenie maszynowe)

NoSQL – No Structured Query Language (nierelacyjna baza danych)

ONZ – Organizacja Narodów Zjednoczonych

PBSSP – Program badań statystycznych statystyki publicznej

PDA – Portable Device Assistant

PKB – produkt krajowy brutto

PKD – Polska Klasyfikacja Działalności

PLS-SEM – Partial Least Squares Structural Equation Modeling

REGON – numer przedsiębiorstwa w rejestrze przedsiębiorstw

SCM – Supply Chain Management

Skills-OVATE – Skills Online Vacancy Analysis Tool for Europe

SMI – Social Media Index

SPC – Statistical Process Control

SQL – Structured Query Language

SVM – Support Vector Machines (uczenie maszynowe)

u.o.b.d. – ustawa z dnia 27 lipca 2001 r. o ochronie baz danych (tekst jedn.: Dz. U. z 2024 r. poz. 1769)

UNECE – United Nations Economic Commission for Europe

URL – Uniform Resource Locator

VR – Virtual Reality

WIN – Web Intelligence Network

WP (ESSnet WP) – Work Package (pakiety robocze grantów Komisji Europejskiej)

Wprowadzenie

Wraz z rozwojem technologii informatycznych ilość danych wzrasta w sposób logarytmiczny. Dane te odzwierciedlają pewien fragment rzeczywistości, a odpowiednio przetworzone dostarczają informacji. Z kolei informacja uzyskana na podstawie danych zwiększa wiedzę odbiorcy o otaczającej go rzeczywistości (Swacha, 2009). Obecnie często obowiązującym terminem opisującym dane zamieszczone w Internecie jest *big data*. Termin ten nie ma jednoznacznej definicji. Początkowo oznaczał duże zbiory danych, później również sposoby pozyskiwania, zapisywania i analizowania informacji, czyli metody *big data* (Brosz, Bryda, Siuda, 2017). Obecnie *big data* jest rdzeniem wielu projektów sztucznej inteligencji (ang. *artificial intelligence*, AI), pozwalającym na dostarczanie coraz lepszej jakościowo informacji wynikowej opierającej się na zbiorach *big data* (Bruchhaus et al., 2025). Jednak najbardziej znana definicja *big data* odnosi się do perspektywy źródła danych i oznacza duży wolumen danych, szybko zmieniający się i cechujący dużą różnorodnością (Laney, 2001). Dane tego rodzaju zamieszczone w Internecie można podzielić na ustrukturyzowane, nieustrukturyzowane oraz częściowo ustrukturyzowane. W obszarze biznesowym dane ustrukturyzowane to np. cenniki produktów dostępne na stronach internetowych czy zestawienia oferowanych produktów. Nieustrukturyzowane obejmują m.in. opinie konsumentów zamieszczane na forach internetowych. Mówiąc o częściowo ustrukturyzowanych, nierzadko ma się na myśli wpisy w mediach społecznościowych, uporządkowane według atrybutów, takich jak data, lokalizacja, nazwa użytkownika, język publikacji, wraz z treścią w postaci nieustrukturyzowanej. Inne często wykorzystywane źródła częściowo ustrukturyzowane obejmują np. strony internetowe z ofertami pracy lub listą oferowanych nieruchomości. Część danych może pozostawać w formie ustrukturyzowanej, np. miejscowość, której dotyczy ogłoszenie, inne pozostają w formie nieustrukturyzowanej, np. szczegółowy opis oferty pracy czy nieruchomości. Tego rodzaju dane mogą być szeroko wykorzystywane w działalności biznesowej, m.in. do zwiększenia konkurencyjności oferowanego asortymentu poprzez proponowanie zniżek na wybrane produkty, w zależności od oferty konkurencji. Inny przykład to wyszukiwanie w Internecie potencjalnych klientów, takich jak firmy z określonej branży lub zajmujące się specyficzną działalnością.

Większość artykułów naukowych koncentruje się na metodach przetwarzania wybranych aspektów stron internetowych. Dotyczy to na przykład działalności e-commerce (García, García-Nieto, Aldana-Montes, 2016).

Przesłanką do napisania monografii było zidentyfikowanie luki badawczej określonej jako brak jednolitej metody, która pozwalałaby na prowadzenie na szeroką skalę badań organizacji na podstawie ich stron internetowych. Pomimo że ekstrakcja danych ze stron internetowych w celu zasilania systemów informatycznych nie jest nowym zjawiskiem i była już przedmiotem wielu badań naukowych (Abramowicz, Kalczyński, Węcel, 2002), to jednak rozwój metod analitycznych, w tym powszechne stosowanie uczenia maszynowego (ang. *machine learning*, ML), znacząco zmienia sposób przetwarzania informacji pochodzącej z Internetu. Metody te pozwalają na pozyskiwanie różnego rodzaju atrybutów organizacji, takich jak rodzaj działalności, poruszane tematy, oferowane produkty, cennik produktów itp. Dostęp do tego rodzaju informacji pozwala między innymi na zwiększenie konkurencyjności jednostki biznesowej poprzez monitorowanie podobnych organizacji, dotarcie do potencjalnych klientów lub firm zainteresowanych współpracą czy też weryfikację aktualnych cen produktów w celu oferowania konkurencyjnych cen. Więcej przykładów praktycznego zastosowania danych pobieranych z Internetu zaprezentowano w kolejnych rozdziałach niniejszej monografii.

Zawarte w literaturze przykłady odnoszą się m.in. do wykorzystania szeregu różnych metod, począwszy od ekstrakcji pewnych słów kluczowych, poprzez wykorzystanie bardziej złożonych rozwiązań, takich jak uczenie maszynowe czy głębokie uczenie (ang. *deep learning*, DL) (Bing, Wong, Lam, 2016). Obecnie pojawiają się rozwiązania pokazujące potencjał ekstrakcji różnego rodzaju informacji ze stron internetowych, m.in. poprzez analizę znajdujących się tam słów kluczowych (Kopp, Orlovskiy, 2022). Artykuły te prezentują jedynie rozwiązania, których testy wykonano na wybranych stronach internetowych. Brakuje w nich jednak kompleksowego i ujednoliconego sposobu przejścia od fazy łączenia różnych metod i źródeł w celu pozyskania listy adresów stron internetowych przedsiębiorstw, poprzez fazę przetwarzania danych, kończąc na udostępnianiu wyników. Nie ma obecnie pozycji literaturowej, która ukazałaby wyniki badania w skali prezentowanej w niniejszej monografii, z wykorzystaniem metod opisanych w niniejszym opracowaniu. Zaproponowana metoda, zaprezentowana szczegółowo w rozdziale szóstym, pozwoliła na usystematyzowanie metod i narzędzi, które należy wykorzystać w procesie pozyskiwania i przetwarzania informacji uzyskanej z Internetu. Co ważne, opisane podejście pozwala na rzetelną ocenę wykorzystanych metod, co umożliwia w późniejszej fazie publikowanie wskaźników dotyczących jakości danych.

Celem niniejszej monografii jest zaprezentowanie autorskiej metody badań dotyczącej pozyskiwania i przetwarzania danych pochodzących z Internetu.

Skoncentrowano się na obszarach związanych z wybranymi atrybutami przedsiębiorstw oraz ogólnie wykorzystaniem *big data* w organizacjach działających w Polsce.

Aby zrealizować cel, postawiono trzy pytania badawcze:

- PB1: Czy strona internetowa jest wystarczającym zasobem informacji pozwalającym na identyfikację rodzaju działalności organizacji, której dotyczy?
- PB2: Czy oferty pracy on-line pozwalają na identyfikację trendów w zakresie wykorzystania technologii *big data* w przedsiębiorstwach?
- PB3: Czy istnieje możliwość pozyskania bazy danych adresów URL w celu wiarygodnego badania polskich przedsiębiorstw?

Do każdego z pytań badawczych przyporządkowano następujące hipotezy:

- H1: Treść strony internetowej wybranej organizacji pozwala na określenie rodzaju jej działalności.
- H2: Badanie ofert pracy pozwala zidentyfikować trendy w zawodach powiązanych z *big data*.
- H3: Obecne źródła dostępne on-line pozwalają na budowę operatu do badań stron internetowych organizacji.

Realizację celu głównego podzielono na etapy. Początkowo scharakteryzowano termin *big data* na podstawie przeglądu literatury. Obejmuje to również zagadnienia związane z metodologią oraz oceną jakości tego rodzaju danych. Oceniając dojrzałość tego typu rozwiązań, dokonano także przeglądu użyteczności ich wykorzystania w nauce i biznesie. Następnie zaprezentowano metody dotyczące przygotowania, pozyskania, a następnie przetwarzania danych pobieranych ze stron internetowych przedsiębiorstw. Przeprowadzone badanie pozwoliło również pokazać wyzwania z tym związane oraz metody zapobiegania problemom pojawiającym się w trakcie przetwarzania tego rodzaju danych.

Realizacja celu głównego monografii nie byłaby możliwa bez przetestowania proponowanego rozwiązania. Aby to osiągnąć, przygotowano zestaw adresów stron internetowych odpowiednio przetworzonych na potrzeby pokazanego badania. Następnie dane pobierano z wielu źródeł, tj. stron internetowych przedsiębiorstw, jak również stron zawierających oferty pracy. Wyniki zostały zaprezentowane zarówno w postaci liczb całkowitych, tj. odnoszących się do liczby jednostek posiadających dane atrybuty, jak również w postaci procentowej, prezentując udział pewnej wielkości w odniesieniu do liczby badanych jednostek.

Użyteczność badania to przede wszystkim jego replikowalność. Każde przedsiębiorstwo posiadające zasoby typu *big data*, tj. oprogramowanie pozwalające na przetwarzanie dużych zbiorów danych, kadre znającą metody pozyskiwania danych z Internetu oraz pracy ze zbiorami *big data*, stosując metody opisane w niniejszej monografii, może zrealizować podobne badanie ukierunkowane na określone informacje, które chce uzyskać. Może to być przydatne między innymi w zakresie poszukiwania partnerów biznesowych, klientów itp. Często bazy danych zawierające

sumaryczne informacje o różnych organizacjach, tzw. *yellow pages*, nie posiadają wystarczająco dużo atrybutów tych jednostek. Prezentowane rozwiązanie pozwala na rozszerzanie zakresu tematycznego badania o dowolne atrybuty, o ile są one możliwe do pozyskania ze stron internetowych.

Niniejsza monografia została podzielona na siedem rozdziałów. W pierwszym rozdziale przedstawiono istotę zbiorów danych i narzędzi typu *big data*. W rozdziale drugim scharakteryzowano źródła *big data* w kontekście typologii danych zawartych w tych źródłach. Trzeci rozdział dotyczy metodologicznych aspektów badań *big data*, scharakteryzowano w nim metody wykorzystywane w badaniach *big data* w kontekście zbiorów danych pochodzących z Internetu. Rozdział czwarty prezentuje problematykę dotyczącą pomiaru jakości danych typu *big data*. W rozdziale piątym dokonano przeglądu literatury w kontekście zastosowania *big data* w nauce, skupiając się na Internecie jako źródle danych *big data*. W rozdziale szóstym zaprezentowano metodykę badania stron internetowych organizacji działających w Polsce. W rozdziale siódmym znajdują się wyniki przeprowadzonego badania. W rozdziale tym przedstawiono również wyzwania związane z prowadzeniem badań na podstawie stron internetowych.

Zrealizowane badanie pozwoliło na uzyskanie odpowiedzi na postawione pytania badawcze, jak również na potwierdzenie sformułowanych hipotez.