

**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