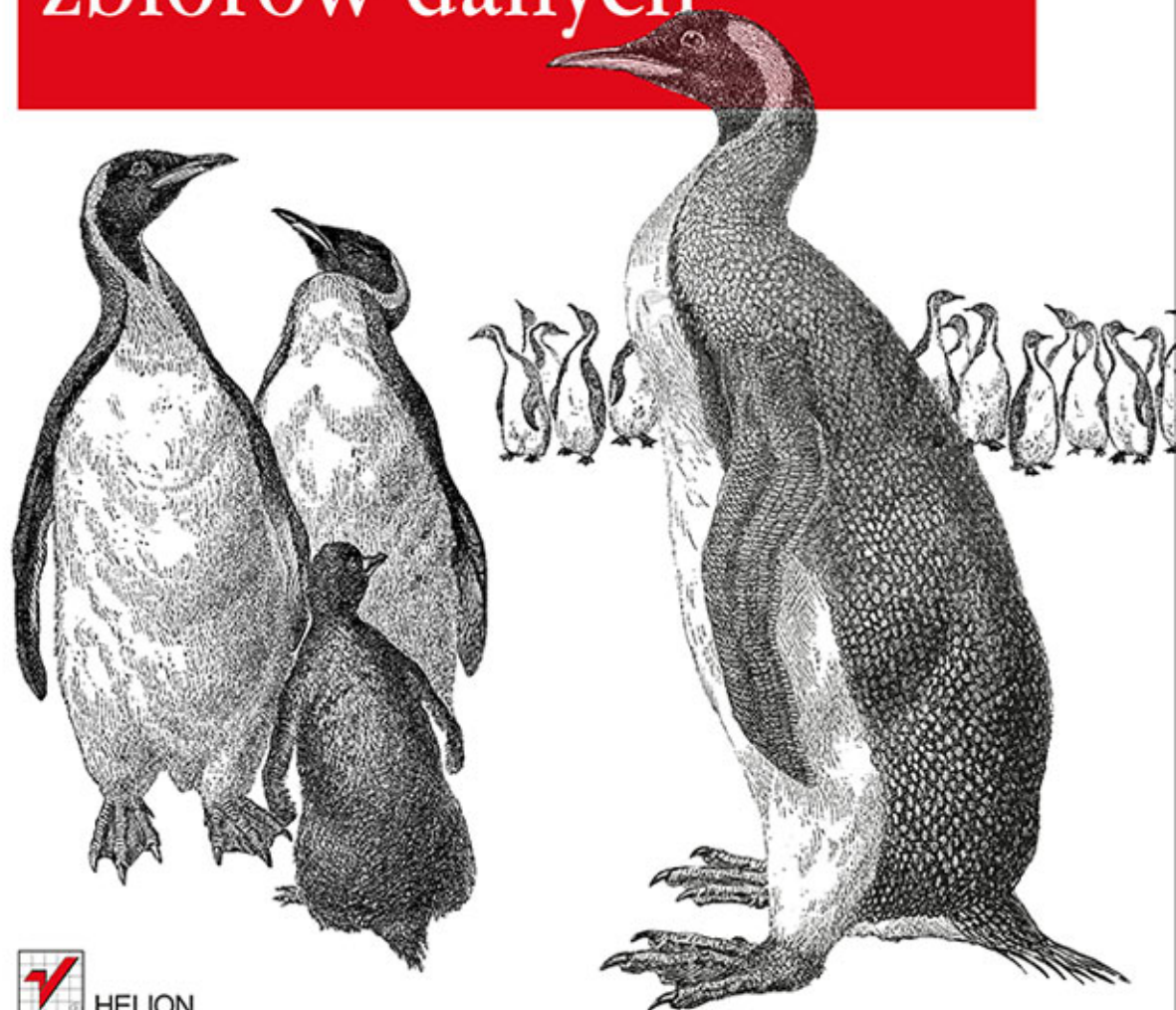


Wykorzystaj dane z sieci do własnych potrzeb!

# Nowe usługi 2.0

## Przewodnik po analizie zbiorów danych



O'REILLY®

*Toby Segaran*

Tytuł oryginału: Programming Collective Intelligence: Building Smart Web 2.0 Applications

Tłumaczenie: Piotr Pilch

ISBN: 978-83-246-9298-9

© 2014 Helion S.A.

Authorized Polish translation of the English edition Programming Collective Intelligence  
ISBN 9780596529321 © 2007 Toby Segaran.

This translation is published and sold by permission of O'Reilly Media, Inc.,  
which owns or controls all rights to publish and sell the same.

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means,  
electronic or mechanical, including photocopying, recording or by any information storage retrieval system,  
without permission from the Publisher.

Wszelkie prawa zastrzeżone. Nieautoryzowane rozpowszechnianie całości lub fragmentu niniejszej  
publikacji w jakiegokolwiek postaci jest zabronione. Wykonywanie kopii metodą kserograficzną,  
fotograficzną, a także kopiowanie książki na nośniku filmowym, magnetycznym lub innym powoduje  
naruszenie praw autorskich niniejszej publikacji.

Wszystkie znaki występujące w tekście są zastrzeżonymi znakami firmowymi bądź towarowymi ich  
właścicieli.

Autor oraz Wydawnictwo HELION dołożyli wszelkich starań, by zawarte w tej książce informacje były  
kompletne i rzetelne. Nie bierze jednak żadnej odpowiedzialności ani za ich wykorzystanie, ani za  
związane z tym ewentualne naruszenie praw patentowych lub autorskich. Wydawnictwo HELION nie  
ponosi również żadnej odpowiedzialności za ewentualne szkody wynikłe z wykorzystania informacji  
zawartych w książce.

Wydawnictwo HELION  
ul. Kościuszki 1c, 44-100 GLIWICE  
tel. 32 231 22 19, 32 230 98 63  
e-mail: helion@helion.pl  
WWW: <http://helion.pl> (księgarnia internetowa, katalog książek)

Pliki z przykładami omawianymi w książce można znaleźć pod adresem:  
<ftp://ftp.helion.pl/przyklady/noweus.zip>

Drogi Czytelniku!  
Jeżeli chcesz ocenić tę książkę, zajrzyj pod adres  
<http://helion.pl/user/opinie/noweus>  
Możesz tam wpisać swoje uwagi, spostrzeżenia, recenzję.

Printed in Poland.

- [Kup książkę](#)
- [Poleć książkę](#)
- [Oceń książkę](#)

- [Księgarnia internetowa](#)
- [Lubię to!» Nasza społeczność](#)

---

# Spis treści

<b>Słowo wstępne .....</b>	<b>11</b>
<b>Przedmowa .....</b>	<b>13</b>
<b>1. Inteligencja zbiorowa — wprowadzenie .....</b>	<b>21</b>
Czym jest inteligencja zbiorowa?	22
Czym jest uczenie maszynowe?	23
Ograniczenia uczenia maszynowego	24
Rzeczywiste przykłady	24
Inne zastosowania algorytmów uczących	25
<b>2. Tworzenie rekomendacji .....</b>	<b>27</b>
Filtrowanie grupowe	27
Gromadzenie preferencji	28
Znajdowanie podobnych użytkowników	29
Rekomendowanie pozycji	34
Dopasowywanie produktów	36
Tworzenie systemu rekomendowania odnośników del.icio.us	38
Filtrowanie oparte na pozycjach	42
Zastosowanie zbioru danych MovieLens	45
Filtrowanie oparte na użytkownikach czy pozycjach?	46
Ćwiczenia	47
<b>3. Wykrywanie grup .....</b>	<b>49</b>
Porównanie uczenia nadzorowanego z nienadzorowanym	49
Wektory wyrazów	50
Grupowanie hierarchiczne	53
Rysowanie dendrogramu	57
Grupowanie kolumn	59
Grupowanie k-średnich	61

Klasyfikacja preferencji	64
Wyświetlanie danych w dwóch wymiarach	68
Inne rzeczy, które mogą być grupowane	71
Ćwiczenia	72
<b>4. Wyszukiwanie i klasyfikowanie .....</b>	<b>73</b>
Co znajduje się w wyszukiwarce?	73
Prosty przeszukiwacz	75
Budowanie indeksu	77
Odpytywanie	81
Klasyfikacja oparta na treści	83
Użycie odnośników zewnętrznych	87
Uczenie na podstawie kliknięć	91
Ćwiczenia	101
<b>5. Optymalizacja .....</b>	<b>103</b>
Podróż grupy osób	104
Reprezentowanie rozwiązań	105
Funkcja kosztu	106
Wyszukiwanie losowe	108
Metoda największego wzrostu	109
Symulowane wyżarzanie	111
Algorytmy genetyczne	113
Wyszukiwania rzeczywistych lotów	117
Optymalizowanie pod kątem preferencji	122
Wizualizacja sieci	125
Inne możliwości	130
Ćwiczenia	130
<b>6. Filtrowanie dokumentów .....</b>	<b>133</b>
Filtrowanie spamu	133
Dokumenty i wyrazy	134
Trenowanie klasyfikatora	135
Obliczanie prawdopodobieństw	137
Naiwny klasyfikator	139
Metoda Fishera	142
Utrwalanie klasyfikatorów po przeprowadzonym treningu	146
Filtrowanie kanałów informacyjnych blogów	148
Poprawianie wykrywania właściwości	150
Użycie interfejsu Akismet	152
Alternatywne metody	153
Ćwiczenia	154

<b>7. Modelowanie przy użyciu drzew decyzyjnych .....</b>	<b>157</b>
Przewidywanie rejestracji	157
Wprowadzenie do drzew decyzyjnych	159
Uczenie drzewa	160
Wybór najlepszego podziału	162
Budowanie drzewa rekurencyjnego	164
Wyświetlanie drzewa	166
Klasyfikowanie nowych obserwacji	168
Przycinanie drzewa	169
Radzenie sobie z brakującymi danymi	171
Radzenie sobie z wynikami liczbowymi	172
Modelowanie cen domów	173
Modelowanie „atrakcyjności”	176
Kiedy stosować drzewa decyzyjne?	178
Ćwiczenia	179
<b>8. Budowanie modelu cen .....</b>	<b>181</b>
Budowanie przykładowego zbioru danych	181
Metoda k-najbliższych sąsiadów	183
Sąsiednie elementy z określoną wagą	186
Walidacja krzyżowa	189
Zmienne heterogeniczne	191
Optymalizowanie skali	194
Rozkłady niejednolite	196
Użycie rzeczywistych danych — interfejs API serwisu eBay	200
Kiedy używać metody k-najbliższych sąsiadów?	207
Ćwiczenia	207
<b>9. Zaawansowane klasyfikowanie: metody jądrowe i maszyny wektorów nośnych .....</b>	<b>209</b>
Zbiór danych swatki	209
Trudności związane z danymi	211
Podstawowa klasyfikacja liniowa	213
Właściwości skategoryzowane	217
Skalowanie danych	218
Metody jądrowe	220
Maszyny wektorów nośnych	223
Zastosowanie biblioteki LIBSVM	225
Dopasowywanie w serwisie Facebook	227
Ćwiczenia	232

<b>10. Znajdowanie niezależnych właściwości .....</b>	<b>233</b>
Zbiór artykułów	234
Wcześniejsze rozwiązania	237
Nieujemna faktoryzacja macierzy	240
Wyświetlanie wyników	246
Użycie danych rynku giełdowego	249
Ćwiczenia	254
<b>11. Inteligencja rozwojowa .....</b>	<b>255</b>
Czym jest programowanie genetyczne?	255
Programy w postaci drzew	258
Tworzenie populacji początkowej	261
Testowanie rozwiązania	263
Krzyżowanie	267
Budowanie środowiska	269
Prosta gra	272
Dalsze możliwości	276
Ćwiczenia	278
<b>12. Algorytmy — podsumowanie .....</b>	<b>281</b>
Klasyfikator bayesowski	281
Klasyfikator drzew decyzyjnych	285
Sieci neuronowe	288
Maszyny wektorów nośnych	292
Metoda k-najbliższych sąsiadów	296
Grupowanie	299
Skalowanie wielowymiarowe	303
Nieujemna faktoryzacja macierzy	305
Optymalizacja	307
<b>A Zewnętrzne biblioteki .....</b>	<b>311</b>
Universal Feed Parser	311
Python Imaging Library	311
Beautiful Soup	312
pysqlite	313
NumPy	314
matplotlib	315
pydelicious	316

<b>B</b>	<b>Formuły matematyczne .....</b>	<b>317</b>
	Odległość euklidesowa	317
	Współczynnik korelacji Pearsona	317
	Średnia ważona	318
	Współczynnik Tanimoto	319
	Prawdopodobieństwo warunkowe	319
	Niejednorodność Giniego	320
	Entropia	321
	Wariancja	321
	Funkcja Gaussa	322
	Iloczyny skalarne	322
	<b>Skorowidz .....</b>	<b>324</b>





# Inteligencja zbiorowa — wprowadzenie

Netflix to internetowa wypożyczalnia płyt DVD, która umożliwia wybór filmów z wysyłką do domu. Firma podaje rekomendacje na podstawie filmów, które zostały wcześniej wypożyczone przez klientów. Pod koniec 2006 r. firma Netflix poinformowała o nagrodzie w wysokości 1 mln dolarów dla pierwszej osoby, która poprawi dokładność systemu rekomendacji wypożyczalni o 10%. Ponadto każdego roku firma będzie wręczać dodatkowo 50 tys. dolarów aktualnemu liderowi do czasu trwania konkursu. W konkursie wzięły udział tysiące zespołów z całego świata. W kwietniu 2007 r. najlepszemu zespołowi udało się uzyskać poprawę rekomendacji o 7%. Korzystając z danych dotyczących filmów, które spodobały się poszczególnym klientom, firma Netflix ma możliwość rekomendowania filmów innym klientom. Ci klienci mogli nawet nigdy o nich nie słyszeć. Po ich obejrzeniu mogą oni zdecydować się na kolejne filmy. Każdy sposób ulepszenia swojego systemu rekomendacji wart jest dla firmy Netflix mnóstwo pieniędzy.

Wyszukiwarka internetowa firmy Google zaczęła działać w 1998 r. W tamtym czasie istniało już kilka dużych wyszukiwarek. Wiele osób przyjęło, że nowy gracz nie będzie w stanie konkurować z gigantami branżowymi. Jednakże założyciele firmy Google zastosowali zupełnie nową metodę tworzenia rankingów wyników wyszukiwania, korzystając z odnośników na milionach stron internetowych podczas określania, które strony są najodpowiedniejsze. Wyniki wyszukiwania wyszukiwarki Google okazały się znacznie lepsze od oferowanych przez innych graczy, którzy w 2004 r. obsługiwali 85% wyszukiwań w Internecie. Założyciele firmy Google zaliczają się obecnie do najbogatszych ludzi świata.

Co te dwie firmy mają ze sobą wspólnego? Obie doszły do nowych wniosków i stworzyły nowe możliwości biznesowe, korzystając z zaawansowanych algorytmów w celu połączenia danych zebranych od wielu różnych osób. Możliwość gromadzenia informacji i moc obliczeniowa pozwalająca na ich interpretowanie stworzyły ogromne możliwości współpracy oraz lepszego zrozumienia użytkowników i klientów. Tego rodzaju działania mają miejsce w różnych przypadkach. Serwisom randkowym zależy na szybszym znalezieniu najlepiej dopasowanych kandydatów. Pojawiają się firmy przewidujące zmiany cen biletów lotniczych. Niemal każdemu zależy na lepszym zrozumieniu swoich klientów, aby przygotować reklamy trafiające do właściwszych osób.

To tylko kilka przykładów ekscytującej dziedziny inteligencji zbiorowej. Rozpowszechnianie się nowych usług powoduje, że każdego dnia pojawiają się nowe możliwości. Wierzę, że opanowanie uczenia maszynowego i metod statystycznych stanie się jeszcze ważniejsze w przeróżnych dziedzinach, szczególnie w przypadku interpretowania i organizowania ogromnej ilości informacji tworzonych przez ludzi na całym świecie.

# Czym jest inteligencja zbiorowa?

Pojęcie *inteligencji zbiorowej* jest używane od dziesięcioleci. Jego znaczenie i popularność zaczęły się zwiększać wraz z pojawieniem się nowych technologii komunikacji. Choć nazwa terminu może przywołać na myśl pojęcia związane ze świadomością zbiorową lub zjawiskiem nadprzyrodzonym, to używając go, specjaliści od technologii mają zwykle na myśli łączenie zachowań, preferencji lub pomysłów grupy osób w celu uzyskania nowatorskich spostrzeżeń.

Oczywiście inteligencja zbiorowa była możliwa przed pojawieniem się Internetu. Nie jest wymagana sieć WWW, aby zgromadzić dane od różnych grup ludzi, połączyć je i poddać analizie. Jedną z najbardziej podstawowych form inteligencji zbiorowej jest ankieta lub spis ludności. Zbieranie odpowiedzi od dużej grupy ludzi umożliwia uzyskanie wniosków statystycznych dotyczących grupy, których nie określiliby w pojedynkę żaden jej członek. Tworzenie nowych wniosków przy udziale niezależnych uczestników w rzeczywistości jest tym, do czego służy inteligencja zbiorowa.

Jej dobrze znanym przykładem są rynki finansowe, w przypadku których cena nie jest ustalana przez jedną osobę ani nie jest wynikiem skoordynowanego działania, lecz stanowi efekt operacji handlowych wielu niezależnych od siebie osób, które działają zgodnie z tym, co w ich przekonaniu służy ich najlepszemu interesowi. Choć z początku może się to wydawać sprzeczne z intuicją, *rynki kontraktów terminowych*, gdzie wielu uczestników handluje kontraktami, próbując określić ich przyszłe ceny, są uważane za lepsze w przewidywaniu cen niż eksperci, którzy niezależnie przygotowują prognozy. Wynika to stąd, że w przypadku tworzenia przewidywań takie rynki łączą w sobie wiedzę, doświadczenie i spostrzeżenia tysięcy osób, a nie analizują jedynie punkt widzenia jednej osoby.

Chociaż metody inteligencji zbiorowej istniały przed powstaniem Internetu, możliwość gromadzenia informacji od tysięcy, a nawet milionów osób w sieci internetowej stworzyła wiele nowych opcji analizy. Ludzie używają Internetu do robienia zakupów, prowadzenia badań, szukania rozrywki i budowania własnych witryn internetowych. Wszystkie te działania mogą być monitorowane i wykorzystywane do uzyskiwania informacji bez żadnej konieczności wpływania na intencje użytkownika przez zadawanie mu pytań. Istnieje ogromna liczba możliwych metod przetwarzania i interpretowania tych informacji. Oto dwa kluczowe przykłady prezentujące przeciwstawne metody.

- *Wikipedia* to internetowa encyklopedia stworzona w całości w wyniku współpracy użytkowników. Dowolna strona może zostać utworzona lub zmodyfikowana przez każdego. Niewielka liczba administratorów monitoruje powtarzające się nadużycia. Serwis *Wikipedia* ma więcej wpisów niż jakakolwiek inna encyklopedia. Pomimo manipulacji dokonywanych przez użytkowników o złych intencjach, generalnie może być uważana za dokładną w przypadku większości zagadnień. Jest to przykład inteligencji zbiorowej, ponieważ każdy artykuł jest utrzymywany przez dużą grupę osób. Efektem jest encyklopedia znacznie większa od jakiegokolwiek, która mogłaby zostać stworzona przez dowolną pojedynczą, skoordynowaną grupę. Oprogramowanie encyklopedii *Wikipedia* nie realizuje żadnych wyszukanych operacji w odniesieniu do uczestniczących użytkowników. Po prostu śledzi zmiany i wyświetla najnowszą wersję.
- Wspomniana wcześniej wyszukiwarka *Google* to najpopularniejsza na świecie wyszukiwarka internetowa. Jako pierwsza zaczęła oceniać strony internetowe na podstawie liczby innych stron, które się do nich odwołują. W przypadku tej metody oceniania uzyskiwane

są informacje o tym, co tysiące osób stwierdziły na temat konkretnej strony internetowej. Informacje te służą do tworzenia rankingu wyników wyszukiwania. Jest to rzykiad inteligencji zbiorowej bardzo odmienny od Wikipedii. Serwis Wikipedia wprost zaprasza swoich użytkowników do uczestnictwa, natomiast wyszukiwarka Google wydobywa ważne informacje o tym, jakie działania twórcy treści sieciowych podejmują w obrębie własnych witryn, a następnie wykorzystuje je do generowania wyników dla swoich użytkowników.

Choć Wikipedia stanowi znakomity zasób i wyjątkowy przykład inteligencji zbiorowej, swoje istnienie zawdzięcza bardziej bazie użytkowników dodających informacje niż sprytnym algorytmom zawartym w oprogramowaniu. W książce skoncentrowano się na drugim końcu tego spektrum, czyli na omówieniu algorytmów takich jak PageRank wyszukiwarki Google, który pobiera dane użytkownika i przeprowadza obliczenia w celu utworzenia nowych informacji mogących wpłynąć na poprawę komfortu obsługi użytkownika. Część danych jest gromadzona jawnie, być może przez prośbę o ocenę różnych rzeczy skierowaną do internautów, a część jest zbierana mimochodem przez obserwację tego, co jest przez nich kupowane. W obu przypadkach istotne jest nie samo zbieranie i wyświetlanie informacji, lecz przetwarzanie ich w inteligentny sposób i generowanie nowych wiadomości.

W książce zostaną zaprezentowane metody gromadzenia danych za pośrednictwem otwartych interfejsów API. Przedstawione będą różne algorytmy uczenia maszynowego oraz metody statystyczne. Taka kombinacja umożliwi przygotowanie metod inteligencji zbiorowej dla danych uzyskanych we własnych aplikacjach, a także gromadzenie danych z innych miejsc i eksperymentowanie z ich wykorzystaniem.

## Czym jest uczenie maszynowe?

Uczenie maszynowe to dziedzina podlegająca sztucznej inteligencji związanej z algorytmami, które umożliwiają *uczenie* komputerów. W większości przypadków oznacza to, że algorytm otrzymuje zbiór danych i określa wnioski dotyczące ich właściwości. Informacje te umożliwiają tworzenie przewidywań odnośnie do innych danych, które mogą pojawić się w przyszłości. Jest to możliwe, ponieważ niemal wszystkie nielosowe dane zawierają wzorce, które pozwalają maszynie dokonywać uogólnień. W tym celu trenowany jest *model* przy użyciu tego, co maszyna uzna za najważniejsze aspekty danych.

Aby zrozumieć, jak powstają modele, rozważmy prosty przykład ze skomplikowanej dziedziny, jaką jest filtrowanie poczty elektronicznej. Załóżmy, że otrzymywana jest spora ilość spamu, który zawiera słowa „apteka internetowa”. Człowiek ma odpowiednie możliwości rozpoznawania wzorców, dlatego potrafi szybko stwierdzić, że każda wiadomość zawierająca te dwa słowa to spam, który powinien trafić bezpośrednio do kosza. Jest to uogólnienie. W rzeczywistości został utworzony myślowy model tego, czym jest spam. Po zgłoszeniu kilku takich wiadomości jako spamu algorytm uczenia maszynowego zaprojektowany do filtrowania spamu powinien być w stanie dokonać takiego samego uogólnienia.

Istnieje wiele różnych algorytmów uczenia maszynowego, cechujących się różną siłą działania i dopasowanych do różnego typu problemów. Niektóre z nich, takie jak drzewa decyzyjne, są transparentne. Oznacza to, że obserwator może w pełni pojąć proces rozumowania realizowany przez maszynę. Inne algorytmy, takie jak sieci neuronowe, to „czarna skrzynka”. Oznacza to, że generują one odpowiedź, często jednak bardzo trudne jest odtworzenie związanej z tym rozumowania.

Wiele algorytmów uczenia maszynowego intensywnie korzysta z matematyki i statystyki. Zgodnie z definicją, którą wcześniej podałem, można nawet stwierdzić, że prosta analiza korelacji i regresja to podstawowe formy uczenia maszynowego. W książce nie założono, że czytelnik ma wiedzę z dziedziny statystyki, dlatego podjąłem się próby objaśnienia zastosowanej statystyki w jak najprostszy sposób.

## Ograniczenia uczenia maszynowego

Uczenie maszynowe nie jest pozbawione wad. Algorytmy mają różne możliwości uogólniania dużych zbiorów wzorców. Wzorec, który nie przypomina żadnego wcześniej napotkanego przez algorytm, z dużym prawdopodobieństwem zostanie niewłaściwie zinterpretowany. Ludzie mogą wykorzystywać rozległe doświadczenie i wiedzę o charakterze kulturowym, a także mają niezwykłą zdolność rozpoznawania podobnych sytuacji podczas podejmowania decyzji dotyczących nowych wiadomości. Z kolei metody uczenia maszynowego mogą jedynie uogólniać na podstawie już napotkanych danych, i to w bardzo ograniczony sposób.

Metoda filtrowania spamu, która zostanie przedstawiona w książce, opiera się na występowaniu słów lub fraz bez względu na ich znaczenie lub na strukturę zdań. Choć teoretycznie możliwe jest zbudowanie algorytmu, który uwzględniałby gramatykę, w praktyce dzieje się to rzadko z powodu wymaganych nakładów nieproporcjonalnie dużych w stosunku do uzyskiwanego ulepszenia algorytmu. Zrozumienie znaczenia słów lub ich powiązania z życiem danej osoby wymagałoby znacznie większej ilości informacji niż ta, do której mogą uzyskać dostęp filtry spamu w swojej obecnej postaci.

Poza tym, choć wszystkie metody uczenia maszynowego różnią się pod tym względem, są podatne na możliwość przesadnego uogólniania. Jak z większością rzeczy w życiu, duże uogólnienia oparte na kilku przykładach rzadko są w pełni dokładne. Z pewnością możliwe jest otrzymanie od znajomego ważnej wiadomości e-mail, która zawiera słowa „apteka internetowa”. W tym przypadku poinstruowano by algorytm, że wiadomość nie jest spamem. W rezultacie algorytm mógłby wywnioskować, że komunikaty od tego konkretnego znajomego są możliwe do zaakceptowania. Natura wielu algorytmów uczenia maszynowego jest taka, że mogą one kontynuować proces uczenia wraz z pojawianiem się nowych informacji.

## Rzeczywiste przykłady

W Internecie istnieje wiele witryn, które obecnie gromadzą dane od wielu różnych osób, a ponadto stosują uczenie maszynowe i metody statystyczne w celu skorzystania z nich. Wyszukiwarka Google to prawdopodobnie największe rozwiązanie (nie tylko używa łączy internetowych do tworzenia rankingu stron, ale nieustannie zbiera informacje dotyczące momentu kliknięcia reklam przez różnych użytkowników), które umożliwia firmie Google bardziej skuteczne kierowanie reklam. W rozdziale 4. zostaną omówione wyszukiwarki internetowe i algorytm Page-Rank, który stanowi istotną część systemu rankingowego wyszukiwarki Google.

Inne przykłady obejmują witryny internetowe z systemami rekomendacji. Witryny takich firm, jak Amazon i Netflix, używają informacji o rzeczach kupionych lub wypożyczonych przez ludzi do określania, jacy internauci lub jakie produkty są do siebie podobne, a następnie tworzenia rekomendacji na podstawie historii zakupów. Inne witryny, takie jak Pandora i Last.fm, stosują oceny użytkowników dotyczące różnych zespołów i piosenek, aby tworzyć tematyczne

stacje radiowe z muzyką, która w opinii ich właścicieli powinna być interesująca. W rozdziale 2. omówiono metody budowania systemów rekomendacji.

*Rynki prognostyczne* to także forma inteligencji zbiorowej. Jednym z najbardziej znanych jest serwis Hollywood Stock Exchange (<http://hsx.com/>), w którym użytkownicy handlują akcjami związanymi z filmami i gwiazdami filmowymi. Możliwe jest kupno lub sprzedaż akcji po aktualnej cenie, jeśli wiadomo, że jej ostateczna cena będzie jedną milionową rzeczywistej kwoty w momencie premiery filmu. Ze względu na to, że cena jest zależna od handlujących akcjami, jej wysokość nie jest ustalana przez żadną konkretną osobę, lecz jako wynik działania grupy. Aktualna cena może być przewidywaniem całej grupy dotyczącym wyniku finansowego filmu po premierze. Przewidywania określone przez serwis Hollywood Stock Exchange są często lepsze od opracowywanych przez poszczególnych ekspertów.

Niektóre serwisy randkowe, takie jak eHarmony, używają informacji zebranych od uczestników do określenia, kto byłby odpowiednim kandydatem. Choć takie firmy utrzymują zwykle stosowane metody dopasowywania osób w tajemnicy, całkiem prawdopodobne jest, że dowolna skuteczna metoda będzie uwzględniać ciągle ponawianie oceny na podstawie tego, czy wybrani kandydaci faktycznie zostali do siebie pomyślnie dopasowani.

## Inne zastosowania algorytmów uczących

Metody opisane w książce nie są nowe. Choć przykłady skupiają się na problemach z inteligencją zbiorową w przypadku zastosowań internetowych, znajomość algorytmów uczenia maszynowego może okazać się pomocna dla twórców oprogramowania w wielu innych dziedzinach. Algorytmy te są szczególnie przydatne w obszarach, w których wykorzystuje się duże zbiory danych przeszukiwane pod kątem interesujących wzorców. Oto przykłady.

### *Biotechnologia*

Postępy w technologii sekwencjonowania i badania przesiewowego spowodowały utworzenie ogromnych zbiorów różnych rodzajów danych, takich jak sekwencje kodu DNA, struktury białek, przesiewy związków chemicznych i ekspresja RNA. Techniki uczenia maszynowego są intensywnie wykorzystywane w przypadku wszystkich tego rodzaju danych. Ma to na celu znalezienie wzorców, które zwiększają stopień zrozumienia procesów biologicznych.

### *Wykrywanie oszustw finansowych*

Firmy obsługujące karty kredytowe nieustannie poszukują nowych sposobów wykrywania nielegalnych transakcji. W związku z tym zastosowały one takie techniki, jak sieci neuronowe i logika indukcyjna, aby weryfikować transakcje i wychwytywać przypadki niewłaściwego użycia.

### *System wizyjny*

Interpretowanie obrazów z kamery wideo do celów wojskowych lub obserwacyjnych to aktywny obszar badań. Wiele technik uczenia maszynowego używanych jest w celu podejmowania próby automatycznego wykrywania intruzów, identyfikowania pojazdów lub rozpoznawania twarzy. Szczególnie interesujące jest zastosowanie technik nienadzorowanych, takich jak *niezależna analiza komponentów*, która umożliwia znajdowanie interesujących właściwości w dużych zbiorach danych.

### *Marketing produktów*

Przez bardzo długi czas zrozumienie demografii i trendów było bardziej formą sztuki niż nauką. Zwiększona w ostatnim czasie możliwość gromadzenia danych od konsumentów zapewniła opcje wykorzystania technik uczenia maszynowego, takich jak grupowanie, aby lepiej zrozumieć naturalne podziały istniejące na rynkach i przygotować precyzyjniejsze przewidywania dotyczące przyszłych trendów.

### *Optymalizacja łańcucha dostaw*

Duże organizacje mogą zaoszczędzić miliony dolarów dzięki efektywnemu funkcjonowaniu ich łańcuchów dostaw i dokładnemu przewidywaniu zapotrzebowania na produkty w różnych obszarach. Liczba możliwych metod tworzenia łańcucha dostaw jest ogromna, tak samo jak liczba czynników, które potencjalnie mogą mieć wpływ na popyt. Optymalizacja i techniki uczenia są często używane do analizowania związanych z tym zbiorów danych.

### *Analiza rynków giełdowych*

Od czasu powstania rynku giełdowego ludzie podejmowali próby wykorzystania matematyki do zarobienia większej ilości pieniędzy. Wraz z coraz większym stopniem zaawansowania uczestników rynku akcji stało się konieczne analizowanie większych zbiorów danych i używanie zaawansowanych technik do wykrywania wzorców.

### *Bezpieczeństwo narodowe*

Ogromna ilość informacji jest gromadzona przez agencje rządowe całego świata. Analiza tych danych wymaga od komputerów wykrywania wzorców i wiązania ich z potencjalnymi zagrożeniami.

To zaledwie kilka przykładów intensywnego wykorzystywania uczenia maszynowego. Z powodu tego, że tendencją jest generowanie większej ilości informacji, prawdopodobnie w większej liczbie dziedzin konieczne będzie wykorzystanie uczenia maszynowego i metod statystycznych, gdy ilość informacji przekroczy ludzkie możliwości zarządzania nimi przy użyciu starych sposobów.

Biorąc pod uwagę, jak dużo nowych informacji udostępnianych jest każdego dnia, oczywiście pojawia się znacznie więcej możliwości. Po poznaniu kilku algorytmów uczenia maszynowego zaczną być zauważalne przeróżne miejsca, w których mogą one zostać wykorzystane.

## A

Akismet, 16, 152  
aktualizowanie multiplikatywne, 244  
algorytm, 23, 281  
  backpropagation,  
    *Patrz:* algorytm wstecznej propagacji błędów  
  CART, *Patrz:* CART  
  filtrowania grupowego,  
    *Patrz:* filtrowanie grupowe genetyczny, 113, 116, 256, 308  
  kNN, *Patrz:* kNN  
  NMF, *Patrz:* macierz faktoryzacja nieujemna  
  PageRank, *Patrz:* PageRank sprzężenia wyprzedzającego, 96  
  syntetyzujący inteligencję zbiorową, 16  
  transparentny, 23  
  uczenia maszynowego,  
    *Patrz:* uczenie maszynowe wstecznej propagacji błędów, 93, 97  
  wybór, 209, 255  
  wyodrębniania właściwości,  
    *Patrz:* dane właściwość wyodrębnianie  
  wyróżniania rdzeni wyrazów, 79  
  wyszukiwania pełnotekstowego, 73  
  zmieniający wagi połączeń między węzłami,  
    *Patrz:* algorytm wstecznej propagacji błędów  
Amazon, 24, 27  
analiza  
  komponentów niezależna, 25  
  korelacji, 24  
  rynków giełdowych, 26

API, 16, 23, 39, 173, 176, 201, 227  
Application Programming Interface,  
  *Patrz:* API

## B

backpropagation, *Patrz:* algorytm wstecznej propagacji błędów  
Bayesa twierdzenie,  
  *Patrz:* twierdzenie Bayesa  
baza danych  
  indeksu pełnotekstowego, 77  
  klient-serwer, 74  
  pysqlite, *Patrz:* pysqlite  
  SQLite, 74, 77, 147  
Beautiful Soup, 64, 75, 312  
biblioteka  
  Beautiful Soup, *Patrz:* Beautiful Soup  
  języka Python, 16  
  LIBSVM, *Patrz:* LIBSVM  
  matplotlib, 198  
  NumPy, *Patrz:* NumPy  
  PIL, 57, 128, 311  
  pydelicious, 316  
  urllib2, 75  
biologia obliczeniowa, 49  
bliskość, 54  
blog, 49, 50, 52  
  filtrowanie, 148

## C

CART, 160  
cena, 181, 205, 206, 207  
  licytacji, 181  
centroid, 62  
Classification and Regression Trees, *Patrz:* CART

## D

dane, *Patrz też:* baza danych, zbiór brakujące, 171  
  demograficzne, 49  
  gromadzenie, 23  
  grupowanie, 49, 50, 54  
  liczbowe, 217  
  macierz, *Patrz:* macierz artykułów  
  nieliniowość, 211  
  przekształcenie w liczby, 217  
  skalowanie, 193, 194, 218, 297  
  optymalizacja, 194  
  transformacja do nowej przestrzeni, 221, 293  
  właściwości wyodrębnianie, 233, 235  
  wzajemna zależność zmiennych, 211  
del.icio.us, 16, 38, 39, 316  
demografia, 26  
dendrogram, 57, 60  
Document Object Model,  
  *Patrz:* DOM  
dokument  
  gromadzenie, 73  
  klasyfikacja, 133  
  tabela, 73, 78  
  XML, 51  
DOM, 118  
domena rozwiązania, 308  
drzewo, 258  
  decyzyjne, 23, 49, 157, 159, 168, 169, 172, 178, 181, 212, 285, 287  
  brakujące dane, 171  
  nadmiernie dopasowane, 169, 170  
  przycinanie, 169, 170  
  rekurencyjne, 164

uczenie, 160, 285  
wady, 287  
wyświetlanie, 166, 167  
zalety, 287  
głębokość, 263  
rekurencyjne, 258  
reprezentacja, 259  
składni, 258  
węzeł, 259  
przechowywania, 277  
wyświetlanie, 261

## E

eBay, 16, 200, 202  
Quick Start Guide, 201  
eHarmony, 25  
elitaryzm, 113  
e-mail  
dystrybucja masowa, 158  
identyfikowanie, 133  
entropia, 163, 164, 170, 321  
przyrost informacji, 164

## F

Facebook, 227  
klucz programisty, 227  
sesja, 228  
znajomy, 229, 230  
faktoryzacja macierzy,  
*Patrz:* macierz faktoryzacja  
filtrowanie  
bayesowskie, 50, 139, 140, 141,  
154, 157, 158, 181, 238, 281,  
283  
wady, 284  
zalety, 284  
grupowe, 28, 42, 46, 47  
poczty elektronicznej, 23  
spam, *Patrz:* spam  
filtrowanie  
Fishera metoda, *Patrz:* metoda  
Fishera  
funkcja  
bazowa radialna, 221, 222  
entropii, *Patrz:* entropia  
Gaussa, 188, 322  
kosztu, 106, 124, 130, 244, 307,  
308  
niejednorodności Giniego,  
*Patrz:* niejednorodność  
Giniego  
odejmowania, 187  
odwrotna, 186

określania wag, 100  
pow, 30  
przydatności, 256  
sigmoidalna, 96  
tangensa hiperbolicznego, 95  
ważona kNN, 189, 322

## G

Gaussa funkcja, *Patrz:* funkcja  
Gaussa  
generacja, 113, 256, 308  
początkowa, 261  
Goldberg David, 28  
Google, 21, 22, 24, 88  
granica decyzyjna, 212  
gromadzenie dokumentów, 73  
GroupLens, 45  
grupowanie, 233, 238  
danych, *Patrz:* dane  
grupowanie  
hierarchiczne, 53, 55, 57, 61,  
299, 300, 302  
kolumn, 59  
k-średnich, 62, 299, 301  
wierszy, 59

## H

hill climbing, *Patrz:* metoda  
największego wzrostu  
hiperpłaszczyzna z  
maksymalnym marginesem, 223  
hodowanie, *Patrz:* krzyżowanie  
Holland John, 116  
Hollywood Stock Exchange, 25  
Hot or Not, 16, 176  
HTML, 79

## I

iloczyn skalarny, 215, 216, 221,  
222, 293, 322  
implementacja referencyjna, 14  
indeks pełnotekstowy, 77, 79, 80  
inteligencja  
rozwojowa, 255  
sztuczna, 23  
w grze, 272  
zbiorowa, 22, 25, 249, 255  
interfejs  
Akismet, *Patrz:* Akismet  
API, *Patrz:* API

## J

Jaccarda współczynnik,  
*Patrz:* współczynnik Jaccarda  
język  
Lisp, 258  
Python, *Patrz:* Python  
XML, *Patrz:* XML

## K

kanal informacyjny  
Atom, 51, 234  
RSS, 51, 234  
filtrowanie, 148  
Kayak, 16, 117, 119  
k-centroid, 62  
klastr, 55, 57  
błąd całkowity, 58  
środek, 62  
wysokość, 58  
klasyfikacja  
odnośników zewnętrznych,  
82, 87, 88, 91  
oparta na treści, 82, 83  
klasyfikator, 135, 146, 153, 209, 238  
bayesowski, 140, 154, 157, 181,  
238, 281, 283  
naiwny, 139, 142, 143, 146  
uczenie, 282  
wady, 284  
zalety, 284  
drzew decyzyjnych,  
*Patrz:* drzewo decyzyjne  
Fishera, *Patrz:* metoda Fishera  
liniowy, 213, 214, 216, 220, 293  
oparty  
na regułach, 133, 134, 141  
na właściwości, 134, 141, 145  
k-Nearest Neighbors, *Patrz:* kNN  
kNN, 183, 185, 196, 207, 296, 298  
wady, 299  
wagi, 186, 189, 196, 298  
zalety, 299  
kodu wcięcie, 15  
korelacja Pearsona, 29, 31, 32, 33,  
35, 54, 66, 317  
krzywa  
dzwonowa, 188, 322  
normalna, *Patrz:* krzywa  
dzwonowa  
krzyżowanie, 256, 258, 267, 308  
k-średnia, 62, 183



## L

Last.fm, 24  
LIBSVM, 225, 226, 295  
linia  
  najlepszego dopasowania, 32  
  podziału, 212, 216, 219, 221,  
  223, 224, 292  
lista, 15  
logika indukcyjna, 25

## M

macierz, 240  
  aktualizacji, 244  
  artykułów, 241, 244  
  wyświetlanie, 247  
  danych, *Patrz:* macierz  
  artykułów  
  faktoryzacja niujemna, 50,  
  240, 241, 242, 243, 249, 251,  
  305, 306  
  mnożenie, 240, 243, 305  
  obserwacji, 251  
  transpozycja, 241, 243  
  wag, 241, 242, 243, 305  
  właściwości, 241, 243, 305  
  wyświetlanie, 246, 252  
maksimum lokalne, 271  
mapa samoorganizująca się, 50  
maszyna wektorów nośnych, 50,  
  181, 209, 223, 224, 225, 226, 231, 292  
  wady, 295  
  zalety, 295  
matplotlib, 315  
metoda  
  Fishera, 142, 145  
  k-najbliższych sąsiadów,  
  *Patrz:* kNN  
  modyfikowania rozwiązań, 113  
  największego wzrostu, 109  
  oparta na wektorach  
  i iloczynach skalarnych, 214  
metryka częstości wyrazów, 84, 85,  
  86  
miara  
  odległości, 302  
  podobieństwa, 29, 31, 32, 33,  
  35, 44, 54, 319  
  stopnia niejednorodności, 320  
  ważona, 34, 186, 189  
MLP, 92  
model  
  myślowy, 23  
  przewidyujący ceny, 181, 186,  
  190, 191, 192, 193, 195, 196,  
  198, 205, 206, 207

Multilayer Perceptron, *Patrz:* MLP  
mutacja, 113, 256, 258, 265, 308

## N

nawigowanie, *Patrz:*  
  przeszukiwanie  
Netflix, 21, 24  
niejednorodność Giniego, 162, 163,  
  320  
NMF, *Patrz:* macierz faktoryzacja  
  nieujemna  
Non-Negative Matrix  
  Factorization, *Patrz:* macierz  
  faktoryzacja niujemna  
NumPy, 242, 314

## O

ocena, 83, 88  
  częstości występowania  
  wyrazów, 83  
  liczbowa, 84  
  lokalizacja w dokumencie, 83  
  normalizacja, 84  
  odległość między wyrazami, 83  
odległość  
  euklidesowa, 29, 33, 35, 54,  
  184, 317  
  Manhattan, 33  
  między wyrazami, 83  
odnośnik zewnętrzny, 82  
optymalizacja, 103, 122, 125, 130,  
  307, 309  
  algorytm genetyczny, *Patrz:*  
  algorytm genetyczny  
  funkcja kosztu, *Patrz:* funkcja  
  kosztu  
  łańcucha dostaw, 26  
  metoda największego  
  wzrostu, *Patrz:* metoda  
  największego wzrostu  
  podróży grupy osób, 104, 117,  
  122  
  presja ewolucyjna,  
  *Patrz:* presja ewolucyjna  
  reprezentowanie rozwiązania,  
  105, 123, 125, 126, 130  
  stochastyczna, 103  
  wyszukiwanie losowe,  
  *Patrz:* wyszukiwanie losowe  
  wyżarzanie symulowane,  
  *Patrz:* wyżarzanie  
  symulowane

## P

Page Larry, 88  
PageRank, 23, 24, 88, 89  
pakiet minidom, 118  
Pandora, 24  
plik HTML, 79  
pocztę elektronicznej filtrowanie,  
  23  
populacja, *Patrz:* generacja  
Porter Stemmer, 79  
prawdopodobieństwo, 319  
  gęstość, 196, 198, 322  
  skumulowane, 198  
  warunkowe, 320  
presja ewolucyjna, 256  
problem koktajlowy, 233  
program  
  krzyżowanie,  
  *Patrz:* krzyżowanie  
  miara sukcesu, 264, 269  
  mutacja, *Patrz:* mutacja  
  reprezentacja drzewa,  
  *Patrz:* drzewo  
programowanie  
  funkcyjne, 14  
  genetyczne, 116, 255, 256, 257  
  funkcje, 276  
  gra, 272, 274, 275  
  pamięć, 277  
  program, *Patrz:* program  
  ranking programów, 271  
  środowisko, 269, 277  
  test, 263, 264  
  obiektywne, 14  
  proceduralne, 14  
przeszukiwanie, 73, 75, 80, 81, 83,  
  84, 85, 86, 87  
przewidywanie liczbowe, 181  
przyrost informacji, 164  
punkt średniej, 213  
pysqlite, 313  
Python, 14, 15  
Python Imaging Library, *Patrz:*  
  biblioteka PIL

## R

regresja, 24  
reguła aktualizowania  
  multiplikatywnego, 244  
rekomendacja, 27, 36  
  odnośników, 38, 41  
  sąsiadów, 41  
  tworzenie, 43

rynek

finansowy, 22, 249, 250  
wolumen obrotów, 249, 250  
kontraktów terminowych, 22  
prognostyczny, 25

## S

serwis randkowy, 25, 176, 209  
sieć neuronowa, 23, 25, 49, 74, 92,  
157, 158, 288, 291  
definicja, 93  
funkcja sigmoidalna, 96  
funkcja tangensa  
hiperbolicznego, 95  
neuron, *Patrz:* sieć neuronowa  
węzeł  
perceptronu  
wielowarstwowego,  
*Patrz:* MLP  
sztuczna, 92, 100, 153  
śledząca kliknięcia, 92  
uczenie, 93, 99, 290  
wady, 292  
warstwa ukryta, *Patrz:*  
warstwa ukryta  
węzeł, 92, 94  
zalety, 292  
skalowanie wielowymiarowe, 68,  
303, 304  
słownik, 15, 39, 40  
identyfikatorów adresów  
URL, 84  
ocen, 84  
zagnieżdżony, 28  
spam, 23, 133  
filtrowanie, 24, 133  
łączenie  
prawdopodobieństw, 139,  
144  
obliczanie  
prawdopodobieństwa,  
137, 138, 139, 140, 141, 142  
oparte na regułach, 133  
uczenie się, 134, 135, 138,  
139, 140, 142, 146, 147  
klasyfikator, *Patrz:* klasyfikator  
WordPress, 152  
SpamBayes, 142  
strona  
internetowa, 75  
ocena, *Patrz:* ocena  
Support Vector Machine, *Patrz:*  
maszyna wektorów nośnych  
SVM, *Patrz:* maszyna wektorów  
nośnych  
system rekomendacji, 24

## Ś

średnia  
punkt, *Patrz:* punkt średniej  
ważona, 189, 318  
świadomość zbiorowa, 22

## T

tabela  
dokumentów, 73, 78  
indeks, 94  
tangens hiperboliczny,  
*Patrz:* funkcja tangensa  
hiperbolicznego  
Tanimoto współczynnik,  
*Patrz:* współczynnik Tanimoto  
Tapestry, 28  
technika nienadzorowana,  
*Patrz:* uczenie nienadzorowane  
tekst odnośników, 91  
transformacja wielomianowa, 293  
trik jądrowy, 221, 224, 293, 294  
twierdzenie Bayesa, 140

## U

uczenie  
maszynowe, 14, 16, 23, 25, 159  
grupowanie, 26  
ograniczenia, 24  
nadzorowane, 49, 153, 233,  
238  
nienadzorowane, 50, 233, 299,  
302  
za pomocą drzew  
decyzyjnych, *Patrz:* drzewo  
decyzyjne  
Universal Feed Parser, 51, 234, 311

## W

walidacja krzyżowa, 189, 206  
wariancja, 321  
warstwa  
ukryta, 92, 94  
zapytań, 92  
wektor, 214, 221, 322  
wiadomość e-mail, *Patrz:* e-mail  
wiersz  
poleceń, 14  
zachęty, 14  
Wikipedia, 22  
witryna społecznościowa, 64, 316  
wizualizacja sieci, 125

wolumen obrotów, 249, 250  
WordPress, 152  
współczynnik  
Jaccarda, 33  
korelacji Pearsona,  
*Patrz:* korelacja Pearsona  
Tanimoto, 66, 319  
tłumienia, 88  
wtyczka SpamBayes,  
*Patrz:* SpamBayes  
wykres punktowy, 211  
wyrażenie listowe, 15, 34  
wyszukiwanie  
losowe, 108, 109  
pełnotekstowe, 73  
wyszukiwarka, 88  
Kayak, *Patrz:* Kayak  
pełnotekstowa, 73  
rejestrwanie kliknięć, 91  
wyżarzanie symulowane, 111,  
128, 244, 308

## X

Xerox PARC, 28  
XML, 14, 51

## Y

Yahoo! Finance, 249, 250

## Z

zakładka, 38  
zapytanie klasyfikowanie, 74  
zbiór  
testowy, 189  
uczący, 189  
zmiennych  
heterogenicznych, 191  
nieistotnych, 192, 193  
zbiór danych  
budowanie, 40, 42  
MovieLens, 45  
Zebo, 64, 65  
Zillow, 173  
zmienna  
heterogeniczna, 191  
nieistotna, 192, 193, 297  
wzajemna zależność, 211  
znacznik, 42  
zupa, 64

# PROGRAM PARTNERSKI

GRUPY WYDAWNICZEJ HELION



1. ZAREJESTRUJ SIĘ
2. PREZENTUJ KSIĄŻKI
3. ZBIERAJ PROWIZJĘ

Zmień swoją stronę WWW  
w działający bankomat!

**Dowiedz się więcej i dołącz już dzisiaj!**

<http://program-partnerski.helion.pl>

GRUPA WYDAWNICZA

 **Helion SA**

## Nowe usługi 2.0

### Przewodnik po analizie zbiorów danych



Internet to nic innego jak gigantyczny zbiór danych. Każdy, kto znajdzie sposób na ich umiejętne wykorzystanie, ma szansę zbudować aplikację, która odniesie światowy sukces. Serwisy randkowe, portale społecznościowe, porównywarki cen – to tylko drobna część serwisów, które możesz wykorzystać przy tworzeniu nowej usługi. Jak analizować dane i wyciągnąć wnioski? Na wiele podobnych pytań odpowiada ta jedyna w swoim rodzaju książka.

W trakcie lektury poznasz najlepsze sposoby filtrowania danych, tworzenia rekomendacji, wykrywania grup oraz wyszukiwania i klasyfikowania. Na kolejnych stronach znajdziesz bogaty zbiór informacji poświęconych algorytmom analizującym dane. Ponadto będziesz mieć możliwość zapoznania się z różnymi sposobami optymalizacji, modelowania przy użyciu drzew decyzyjnych oraz tworzenia modeli cenowych. Książka ta w rękach wprawnego programisty może stanowić niesamowite narzędzie. Otwiera wrota do świata pełnego danych i zależności pomiędzy nimi!

Dzięki tej książce:

- poznasz najlepsze i najskuteczniejsze algorytmy do analizy danych,
- zbudujesz model cen,
- nauczysz się korzystać z drzew decyzyjnych,
- zastosujesz dane z sieci do budowy nowych usług.

*Wyciągnij właściwe wnioski z posiadanych danych!*

**helion.pl**  
księgarnia  
internetowa

Nr katalogowy: 24398



Księgarnia internetowa:  
<http://helion.pl>



Zamówienia telefoniczne:  
**0 801 339900**  
**0 601 339900**



**Helion**

Sprawdź najnowsze promocje:

🔗 <http://helion.pl/promocje>

Książki najchętniej czytane:

🔗 <http://helion.pl/bestsellery>

Zamów informacje o nowościach:

🔗 <http://helion.pl/nowosci>

**Helion SA**

ul. Kościuszki 1c, 44-100 Gliwice

tel.: 32 230 98 63

e-mail: [helion@helion.pl](mailto:helion@helion.pl)

<http://helion.pl>



ISBN 978-83-246-9298-9



Cena 54,00 zł