

Inteligentne wydobywanie informacji z internetowych serwisów społecznościowych

AUTOMATYKA INFORMATYKA

Technologie Informacyjne

Sieć Semantyczna

Przetwarzanie Języka Naturalnego

Internet

Edytor Serii: Zdzisław Kowalczyk

Inteligentne wydobywanie informacji z internetowych serwisów społecznościowych

Redakcja:

Bogdan Wiszniewski



Redaktor Naczelny i Edytor Serii: *Zdzisław Kowalczuk, prof. dr hab. inż.*
Redaktor wydania: *Bogdan Wiszniewski, prof. dr hab. inż.*
Korekta wydawnicza: *Anna Osadowska*
Projekt graficzny: *Monika Wiszniewska (Kowalczuk), mgr inż. arch.*
Praca naukowa finansowana przez MNiSzW i NCBR
jako projekt rozwojowy
ze środków na naukę w latach 2009 – 2011
Książka wydana nakładem Towarzystwa Konsultantów Polskich, Oddział Gdańsk

Książkę wydrukowano na podstawie materiałów przygotowanych przez Autorów

Copyright© by TKP, Pomorskie Wydawnictwo Naukowo-Techniczne, Gdańsk 2011

Copyright© by Bogdan Wiszniewski, 2011

All rights reserved

Wszystkie nazwy produktów wymienione w niniejszej publikacji są zastrzeżonymi nazwami handlowymi lub znakami towarowymi odpowiednich firm.

Niniejszej książki w całości lub części nie wolno powielać, ani przekazywać w żaden sposób (nawet za pomocą nośników mechanicznych, elektronicznych i magnetycznych), w tym też umieszczać lub rozpowszechniać w postaci cyfrowej zarówno w Internecie, jak i sieciach lokalnych, bez uzyskania pisemnej zgody Wydawnictwa PWNT Towarzystwa Konsultantów Polskich.

Pomorskie Wydawnictwo Naukowo-Techniczne [PWNT](#)
Towarzystwa Konsultantów Polskich Oddział Gdańsk
80-309 Gdańsk, ul. Grunwaldzka 311
tel./fax 58 552 1536
e-mail: tkp@konsulting.gda.pl
strony: <http://www.konsulting.gda.pl/pwnt> (sklep internetowy)

ISBN 978-83-926806-6-6

Spis treści

Przedmowa

Rozdział 1: Języki i gramatyki formalne	1
1.1 Języki formalne, języki programowania i metajęzyki	1
1.1.1 Języki programowania	2
1.1.2 Translatory, kompilatory oraz interpretery	4
1.2 Gramatyki formalne	4
1.2.1 Definicja i rodzaje gramatyk formalnych	5
1.2.2 Notacje zapisu gramatyk formalnych	6
1.2.3 Klasyfikacja Chomsky'ego	9
1.2.4 Wyrażenia regularne	11
1.2.5 Jednoznaczność, rozstrzygalność i przydatność gramatyk	13
1.3 Analiza zdań języka	14
Rozdział 2: Automaty jako narzędzia w przetwarzaniu języka	17
2.1 Stany automatu i notacja grafowa	17
2.2 Klasyfikacje automatów	19
2.2.1 Funkcja automatu	19
2.2.2 Akceptowane klasy języków	19
2.2.3 Determinizm działania automatu	20
2.3 Automaty skończone	20
2.3.1 Deterministyczne automaty skończone (DFA)	22
2.3.2 Niedeterministyczne automaty skończone (NFA)	22
2.3.3 Niedeterministyczne automaty skończone z ϵ -przejściami (ϵ -NFA)	24
2.4 Tworzenie automatu na podstawie wyrażenia regularnego	24
2.5 Determinizacja automatu	25
2.6 Minimalizacja automatu	27
2.7 Automat Mealy'ego	29
2.8 Determinizacja automatu Mealy'ego	30
Rozdział 3: Przetwarzanie języka naturalnego	33
3.1 Perspektywy przetwarzania języka naturalnego	34
3.2 Zastosowania przetwarzania języka naturalnego	35
3.3 Model warstwowy przetwarzania	36
3.3.1 Warstwa segmentacji	36
3.3.2 Warstwa słownikowa	37
3.3.3 Płytkowa analiza składniowa	39
3.3.4 Warstwa składniowa	40
3.3.5 Warstwa znaczeniowa	42
3.3.6 Warstwa użycia	44

3.4	Jakość przetwarzania	44
3.4.1	Miary jakości	45
3.5	Poprawianie pisowni	45
3.6	Optymalizacja reguł	48
Rozdział 4: Głęboka analiza tekstu w języku polskim		51
4.1	Pragmatyka i funkcje tekstów	52
4.2	Problem homonimii	53
4.3	Analizatory i wyszukiwarki	54
4.4	Dehomonimizacja i desynkretyzacja	55
4.5	Typy analiz	56
4.5.1	Analiza morfologiczna	56
4.5.2	Analiza składniowa	57
4.5.3	Analiza głęboka	58
4.5.4	Przykładowa analiza zdania	58
4.6	Analiza semantyczna	60
4.7	Nowe zjawiska językowe	60
Rozdział 5: Metody wspomaganie wyszukiwania informacji		63
5.1	Języki zapytań wyszukiwarek internetowych	63
5.1.1	Język zapytań w wyszukiwarce Google	64
5.1.2	Język zapytań w innych wyszukiwarkach	64
5.2	Interaktywne rozszerzanie zapytań	65
5.2.1	Metody globalne	66
5.2.2	Metody lokalne	67
5.2.3	Tworzenie rankingu użyteczności słów	67
5.3	Zapytania powiązane tematycznie	67
5.4	Personalizacja	69
5.5	Wykorzystanie modelu przestrzeni wektorowej do wspomaganie wyszukiwania	69
5.6	Wyszukiwanie relewantne	71
5.7	Wyszukiwanie zasobów podobnych	71
5.8	Klasteryzacja dokumentów	72
5.9	Zapytania w języku naturalnym	72
5.10	Chmury znaczników	73
5.11	Metoda klasteryzacji kierunkowej	74
Rozdział 6: Detekcja obiektów graficznych i ekstrakcja ich parametrów		75
6.1	Analiza obrysu obiektu	75
6.1.1	Podział linii brzegowej na tokeny	76
6.1.2	Wykorzystanie symetrii do porównywania kształtów	77
6.2	Analiza zawartości obrazu (tekstury)	79
6.2.1	$N \times M$ -gramy	79
6.2.2	Lokalne wzorce	80
6.2.3	Filtry Gabora	82
6.3	Wykrywanie obiektów metodą AdaBoost	84
6.3.1	Wyznaczanie cech	85
6.3.2	Funkcje klasyfikujące	87
6.3.3	Kaskada klasyfikatorów	88

Rozdział 7: Selekcja i ekstrakcja cech	91
7.1 Reprezentacja danych	91
7.2 Selekcja cech	92
7.2.1 Generowanie podzbioru cech	93
7.2.2 Ocena jakości podzbioru cech	96
7.2.3 Kryterium stopu	99
7.3 Ekstrakcja cech	99
7.3.1 Analiza głównych składowych	100
7.3.2 Wielowymiarowe skalowanie	101
7.3.3 Liniowa analiza dyskryminacyjna	102
Rozdział 8: Algorytmy klasyfikacji i uczenia w rozpoznawaniu treści	105
8.1 Uczenie a uogólnianie	106
8.2 Klasyfikator bayesowski	107
8.2.1 Estymacja parametrów rozkładu normalnego	108
8.2.2 Naiwny klasyfikator bayesowski	110
8.3 Klasyfikator najbliższego sąsiada	110
8.4 Drzewa decyzyjne	110
8.4.1 Algorytm ID3	111
8.4.2 Metody poprawy uogólniania	113
8.5 Sztuczne sieci neuronowe (SSN)	114
8.5.1 Uczenie klasyfikatora neuronowego w oparciu o zbiór przykładów	116
8.5.2 Algorytm propagacji wstecznej błędu	117
8.6 Metoda wektorów wspierających (SVM)	117
Rozdział 9: Studium przypadku – system ISPAD	121
9.1 Moduł pozyskiwania danych	122
9.2 Moduł zarządzania korpusem	123
9.2.1 Narzędzia do pracy nad tekstami o określonych treściach	123
9.2.2 Testy korpusu ISPAD	124
9.3 Moduł analizy danych	125
9.3.1 Normalizacja i segmentacja tekstu	127
9.3.2 Analiza językowa	128
9.3.3 Wnioskowanie	132
9.3.4 Przygotowanie reakcji	136
9.4 Moduł zarządzania bazą wiedzy	137
Bibliografia	139
Skorowidz	145
Streszczenie w jęz. angielskim	149

