

OBSAH

1	ÚVOD DO ANALÝZY INFORMAČNÍCH ZDROJŮ	14
1.1	Základní pojmy	14
1.2	Motto studijních textů.....	15
2	ANALÝZA STRUKTUROVANÝCH DAT	16
2.1	Základní pojmy	16
2.2	Informační systémy	17
2.2.1	Dělení informačních systémů podle úrovně jejich nasazení	17
2.2.2	Architektura informačních systémů.....	18
2.2.3	Životní cyklus informačního systému.....	20
2.3	Databázové systémy	21
2.3.1	Složení databázového systému	22
2.3.2	Schéma databáze	23
2.3.3	Operace nad databází	31
2.3.4	Funkce databázového systému.....	40
2.3.5	Příklady struktury databáze.....	54
2.4	Jazyk SQL.....	58
2.4.1	Princip zpracování SQL dotazů a jejich kategorie.....	58
2.4.2	Data Definition Language.....	60
2.4.3	Data Manipulation Language.....	63
2.4.4	Data Control Language	68
2.4.5	Uložené procedury a spouštěče.....	69
2.4.6	Aplikace jazyka SQL v programovacích jazycích.....	70
2.5	Algoritmy analýzy strukturovaných dat	72
2.5.1	Hashovací funkce.....	72
2.5.2	Třídící algoritmy	75
2.5.3	Vyhledávací algoritmy	80
2.6	Softwarové nástroje	84
2.6.1	PostgreSQL	84
2.6.2	MySQL	85
2.6.3	Oracle Database	86
2.6.4	IBM DB2 a Informix	86
2.6.5	Microsoft SQL Server a Access.....	86

3	ANALÝZA NESTRUKTUROVANÝCH DAT	88
3.1	Základní pojmy	88
3.2	Systémy pro správu dokumentů.....	90
3.2.1	Dokument, jeho vlastnosti a životní cyklus	90
3.2.2	Funkce systémů pro správu dokumentů.....	96
3.3	Dokumentografické informační systémy	103
3.3.1	Architektura dokumentografického informačního systému.....	103
3.3.2	Dokazovací jazyky v dokumentografických informačních systémech.....	105
3.3.3	Moderní možnosti fulltextového vyhledávání	107
3.3.4	Možnosti analýzy obsahu dokumentů.....	109
3.4	Fulltextové vyhledávání.....	112
3.4.1	Vývoj fulltextového vyhledávání.....	112
3.4.2	Proces vyhledávání	114
3.4.3	Proces indexování	116
3.4.4	Hodnocení kvality vyhledávání	134
3.4.5	Možnosti fulltextového vyhledávání u jiných typů dokumentů	137
3.5	Dotazovací jazyk systému Tovek Tools	139
3.5.1	Prvky dotazu	140
3.5.2	Způsoby zápisu dotazu.....	140
3.5.3	Typy operátorů.....	141
3.6	Algoritmy analýzy nestrukturovaných dat.....	148
3.6.1	Algoritmy pro přesné vyhledávání vzorů v textu	148
3.6.2	Algoritmy pro detekci a opravu chyb	152
3.6.3	Algoritmus výběru vhodných výrazů pro indexaci.....	155
3.7	Softwarové nástroje	156
3.7.1	Microsoft SharePoint	156
3.7.2	Tovek Tools	158
3.7.3	i2 Analyst's Notebook	159
3.7.4	Autonomy Virage MediaBin.....	160
4	DATA MINING.....	161
4.1	Základní pojmy	161
4.2	Úvod do data miningu	162
4.2.1	Účel data miningu	163

4.2.2	Historie a vývoj data miningu.....	163
4.2.3	Proces získávání znalostí	164
4.2.4	Proces řešení manažerského problému	168
4.2.5	Vstupní data	169
4.2.6	Typy a formy výstupů.....	171
4.2.7	Ilustrační příklady typických úloh data miningu	178
4.2.8	Aplikace data miningu v praxi.....	192
4.2.9	Data mining a etika	195
4.3	Základní metody data miningu	196
4.3.1	1R algoritmus.....	196
4.3.2	Naivní Bayesův klasifikátor.....	199
4.3.3	Konstrukce rozhodovacích stromů	201
4.3.4	Konstrukce rozhodovacích pravidel	206
4.3.5	Lineární modely	209
4.3.6	Klasifikace podle vzdálenosti	211
4.3.7	Shlukování	212
4.3.8	Vícezáznamová klasifikace.....	214
4.3.9	Kombinace několika metod data miningu	214
4.4	Metody hodnocení a ověření výsledků	215
4.4.1	Hodnocení predikce třídy u klasifikačních úloh	216
4.4.2	Hodnocení přesnosti odhadů pravděpodobností	219
4.4.3	Hodnocení predikce hodnot numerických atributů.....	220
4.4.4	Analýza nákladů.....	220
4.5	Oblasti využití data miningu v AČR	226
4.5.1	Využití data miningu při analýze signálů	226
4.5.2	Využití data miningu při analýze hrozeb a bezpečnostního prostředí	227
4.5.3	Využití data miningu v systémech včasného varování	227
4.6	Softwarové nástroje	230
	LITERATURA.....	234

SEZNAM OBRÁZKŮ

Obr. 2.1	Informační pyramida.....	18
Obr. 2.2	Architektura klient/server.....	19
Obr. 2.3	Architektura orientovaná na služby	20
Obr. 2.4	Typická struktura informačního systému.....	22
Obr. 2.5	Složení databázového systému.....	22
Obr. 2.6	Koncept tabulky v relačním datovém modelu	23
Obr. 2.7	Příklad tabulky	24
Obr. 2.8	Znázornění struktury tabulky	24
Obr. 2.9	Příklady vztahů mezi tabulkami.....	28
Obr. 2.10	Realizace vztahu 1:1	29
Obr. 2.11	Realizace vztahu 1:N	30
Obr. 2.12	Realizace vztahu N:M.....	30
Obr. 2.13	Modelový příklad aplikace vztahové tabulky	31
Obr. 2.14	Modelový příklad databáze v relačním datovém modelu	32
Obr. 2.15	Modelový příklad pro výběrové operace	34
Obr. 2.16	Výsledky operace projekce (vlevo) a restrikce (vpravo)	35
Obr. 2.17	Výsledky operace union (nahore) a join (dole).....	35
Obr. 2.18	Výsledky kombinace operátorů join a projekce.....	36
Obr. 2.19	Výsledek operace join na tabulkách ve vztahu N:M.....	37
Obr. 2.20	Výsledek operací SUM a COUNT.....	38
Obr. 2.21	Modelový příklad tabulky pro vyhledávání	41
Obr. 2.22	Příklady s výsledky jednoduchého vyhledávání	41
Obr. 2.23	Příklady s výsledky porovnávání textových řetězců	42
Obr. 2.24	Základní možnosti regulárních výrazů při tvorbě vzoru	43
Obr. 2.25	Příklady použití zástupných znaků při vyhledávání v řetězcích	43
Obr. 2.26	Princip booleovských operátorů.....	44
Obr. 2.27	Příklady s výsledky dotazů s booleovskými operátory	45
Obr. 2.28	Spojení více podmínek booleovskými operátory a použití závorek	46
Obr. 2.29	Princip algoritmu binárního vyhledávání.....	47
Obr. 2.30	Princip indexace dat	49
Obr. 2.31	Kontrola integrity dat pomocí kontrolního čísla	52
Obr. 2.32	Příklad struktury databáze (vedení osob, událostí a telefonních hovorů)	55

Obr. 2.33	Příklad struktury databáze (knihovna dokumentů)	56
Obr. 2.34	Hlavní dialogové okno knihovního systému.....	57
Obr. 2.35	Fáze zpracování SQL příkazu systémem řízení báze dat.....	59
Obr. 2.36	Princip přístupu k SŘBD prostřednictvím nativních protokolů.....	71
Obr. 2.37	Princip přístupu k SŘBD přes univerzální rozhraní.....	71
Obr. 2.38	Matematická operace XOR.....	72
Obr. 2.39	Hashovací algoritmus modulo	73
Obr. 2.40	Hashovací algoritmus CRC.....	75
Obr. 2.41	Příklad výpočtu 4bitového CRC	75
Obr. 2.42	Algoritmus bublinkového třídění	76
Obr. 2.43	Příklad seřazení množiny prvků bublinkovým algoritmem	77
Obr. 2.44	Koncept algoritmu rychlého třídění (quicksort).....	78
Obr. 2.45	Pravidla při rozdělení množiny u algoritmu rychlého třídění	78
Obr. 2.46	Algoritmus rychlého třídění (quicksort).....	79
Obr. 2.47	Příklad seřazení pole algoritmem rychlého třídění	79
Obr. 2.48	Algoritmus lineárního vyhledávání.....	81
Obr. 2.49	Algoritmus binárního vyhledávání.....	81
Obr. 2.50	Základní princip hashování	82
Obr. 2.51	Příklad vyhledávání hashováním	83
Obr. 2.52	Grafické uživatelské prostředí pgAdmin systému PostgreSQL.....	85
Obr. 3.1	Analogie pojmů strukturovaných a nestrukturovaných dat	89
Obr. 3.2	Ukázka životního cyklu dokumentu	92
Obr. 3.3	Příklad členění kategorií dokumentů	97
Obr. 3.4	Příklad procesu workflow	98
Obr. 3.5	Architektura dokumentografického informačního systému.....	104
Obr. 3.6	Přístup uživatelů k DIS prostřednictvím architektury klient/server.....	105
Obr. 3.7	Rozšířené vyhledávání Google	106
Obr. 3.8	Množina relevantních dokumentů v informačním zdroji.....	107
Obr. 3.9	Zobrazení vazeb mezi dotazy a dokumenty	110
Obr. 3.10	Zobrazení vazeb mezi klíčovými slovy v dokumentech	111
Obr. 3.11	Vývoj fulltextového vyhledávání.....	114
Obr. 3.12	Proces vyhledávání v dokumentografickém systému	115
Obr. 3.13	Proces vyhledávání bez využití principu indexování.....	116
Obr. 3.14	Proces vyhledávání s využitím principu indexování.....	117

Obr. 3.15	Proces při vkládání nového dokumentu do systému	118
Obr. 3.16	Princip invertovaného souboru	119
Obr. 3.17	Příklad invertovaného souboru	120
Obr. 3.18	Rozložení indexu do několika indexových tabulek	121
Obr. 3.19	Příklad rozložení indexu do několika indexových tabulek	121
Obr. 3.20	Rozložení indexu do indexových tabulek podle prvních dvou písmen	122
Obr. 3.21	Rozšíření indexu o další atributy	123
Obr. 3.22	Princip vyhledávání prostřednictvím signaturového souboru.....	125
Obr. 3.23	Realizace signaturového souboru formou tabulek databáze	126
Obr. 3.24	Příklady porovnání signatury dotazu se signaturami dokumentů	126
Obr. 3.25	Příklad reprezentace pojmů jedním bitem	127
Obr. 3.26	Příklad reprezentace pojmů více bity	128
Obr. 3.27	Problém ztráty informace v signatuře při reprezentaci pojmu více bity	128
Obr. 3.28	Modelový příklad dokumentů v informačním zdroji	131
Obr. 3.29	Kosinová míra jako vzdálenost mezi vektorem dotazu a dokumentu.....	133
Obr. 3.30	Množiny dokumentů k hodnocení kvality vyhledávání	134
Obr. 3.31	Nepřímá úměra mezi koeficienty přesnosti a úplnosti.....	136
Obr. 3.32	Vývoj vyhledávání při upřesňování dotazu.....	136
Obr. 3.33	Hodnocení kvality vyhledávače u společnosti Google	137
Obr. 3.34	Konstrukční prvky dotazu v jazyce systému Tovek Tools	140
Obr. 3.35	Hierarchie konceptuálních operátorů BEST, AND a OR	144
Obr. 3.36	Algoritmus vyhledávání vzorů v textu.....	149
Obr. 3.37	Ukázka činnosti algoritmu vyhledávání vzorů v textu.....	149
Obr. 3.38	Tvorba matice X pro SHIFT-OR algoritmus	150
Obr. 3.39	Vektory $Aznak$ pro SHIFT-OR algoritmus.....	151
Obr. 3.40	Vektor R pro SHIFT-OR algoritmus	151
Obr. 3.41	Algoritmus SHIFT-OR	151
Obr. 3.42	Ukázka vývoje vektoru R v průběhu algoritmu	152
Obr. 3.43	Algoritmus detekce a opravy chyb v textovém dokumentu.....	153
Obr. 3.44	Vhodnost výrazů k indexaci v závislosti na jejich výskytu v dokumentech.....	155
Obr. 3.45	Algoritmus výběru vhodných výrazů pro indexaci.....	156
Obr. 3.46	Intranet Univerzity obrany	157
Obr. 3.47	Týmový web Katedry taktiky	158
Obr. 3.48	Ukázka nástroje Tovek Tools.....	159

Obr. 3.49 Ukázka nástroje i2 Analyst's Notebook.....	160
Obr. 4.1 Proces získávání znalostí z dat.....	164
Obr. 4.2 Vytvoření modelu nad metodou data miningu.....	168
Obr. 4.3 Data mining jako prostředek při řešení manažerského problému.....	168
Obr. 4.4 Struktura rozhodovacích stromů.....	173
Obr. 4.5 Jednoduché podmínky v rozhodovacích stromech.....	174
Obr. 4.6 Predikce nominálních nebo numerických výstupních atributů.....	174
Obr. 4.7 Struktura rozhodovacího pravidla.....	175
Obr. 4.8 Příklad seznamu rozhodujících pravidel.....	175
Obr. 4.9 Úprava rozhodujících pravidel z obrázku 4.8 beze změny významu.....	175
Obr. 4.10 Rozhodující seznam.....	176
Obr. 4.11 Rozhodující pravidla s výjimkami.....	176
Obr. 4.12 Asociační pravidlo.....	176
Obr. 4.13 Rozhodující pravidla vzájemně porovnávající hodnoty atributů.....	177
Obr. 4.14 Shlukování vstupních záznamů do skupin.....	177
Obr. 4.15 Vstupní data k úloze s kontaktními ččkami.....	179
Obr. 4.16 Znalost ve formě rozhodovacího pravidla.....	180
Obr. 4.17 Rozšíření rozhodovacího pravidla o další prvek.....	180
Obr. 4.18 Úplný seznam rozhodovacích pravidel pro úlohu s kontaktními ččkami.....	180
Obr. 4.19 Rozhodovací strom pro úlohu s kontaktními ččkami.....	181
Obr. 4.20 Vstupní data k úloze s počasím.....	182
Obr. 4.21 Rozhodovací seznam pro úlohu s počasím.....	183
Obr. 4.22 Vstupní data s numerickými atributy k úloze s počasím.....	184
Obr. 4.23 Rozhodovací pravidlo pro numerický atribut v úloze s počasím.....	185
Obr. 4.24 Asociační pravidla v úloze s počasím.....	185
Obr. 4.25 Vstupní data k úloze klasifikace kosatců.....	186
Obr. 4.26 Rozhodovací pravidla v úloze klasifikace kosatců.....	186
Obr. 4.27 Vstupní data k úloze s odhadem výkonu počítače.....	187
Obr. 4.28 Princip lineárního modelu pro binární klasifikační problém.....	188
Obr. 4.29 Vstupní data k úloze s vyjednáváním odborů.....	189
Obr. 4.30 Rozhodovací strom pro úlohu s vyjednáváním odborů (první varianta).....	190
Obr. 4.31 Rozhodovací strom pro úlohu s vyjednáváním odborů (druhá varianta).....	191
Obr. 4.32 Rozhodovací pravidlo v úloze s klasifikací onemocnění rostlin.....	192
Obr. 4.33 1R algoritmus.....	197

Obr. 4.34	Princip činnosti 1R algoritmu	197
Obr. 4.35	Rozhodovací pravidla vytvořená algoritmem 1R pro úlohu s počasím	197
Obr. 4.36	Seřazení hodnot numerického atributu společně s třídou	198
Obr. 4.37	Rozhodovací pravidla vytvořená algoritmem 1R na numerickém atributu	198
Obr. 4.38	Rozhodovací pravidlo vzniklé spojením intervalu u numerického atributu	199
Obr. 4.39	Stanovení pravděpodobnosti výskytu třídy u úlohy s počasím	199
Obr. 4.40	Příklad výpočtu pravděpodobnosti příslušnosti záznamu ke třídě	200
Obr. 4.41	Výpočet průměru a směrodatné odchylky u numerických atributů	200
Obr. 4.42	Algoritmus pro tvorbu rozhodovacích stromů	202
Obr. 4.43	Volba atributu při konstrukci uzlu rozhodovacího stromu	202
Obr. 4.44	Volba atributu při konstrukci uzlu druhé úrovně rozhodovacího stromu	204
Obr. 4.45	Kompletní rozhodovací strom pro úlohu s počasím	205
Obr. 4.46	Situace při pokrývání množiny vstupních záznamů pravidlem	206
Obr. 4.47	Algoritmus pro tvorbu rozhodovacích pravidel	207
Obr. 4.48	Koncept perceptronu	210
Obr. 4.49	Algoritmus pro stanovení vah perceptronu	210
Obr. 4.50	Algoritmus klasifikace vyhledáním nejbližšího souseda	212
Obr. 4.51	Algoritmus iterativního shlukování	213
Obr. 4.52	Stavy při binární klasifikaci	221
Obr. 4.53	Způsob definice stavů při analýze nákladů pro tři třídy	221
Obr. 4.54	Matice nákladů	222
Obr. 4.55	Graf navýšení	223
Obr. 4.56	Matice nákladů pro úlohu s reklamními letáky	224
Obr. 4.57	Grafy nákladů pro úlohu s reklamními letáky	224
Obr. 4.58	Příklad ROC křivky	225
Obr. 4.59	Význam ROC křivek při porovnávání dvou klasifikačních metod	226
Obr. 4.60	Úspěšnost výpočetních metod při předpovědi konfliktů mezi státy	229
Obr. 4.61	Úspěšnost předpovědi konfliktů mezi státy pomocí metod data miningu	230
Obr. 4.62	Rozhodovací strom pro úlohu předvídání konfliktů vytvořený metodou J48	231
Obr. 4.63	Dialogové okno hlavního uživatelského rozhraní nástroje WEKA	232