

Obsah

1	Úvod	9
1.1	Cíle knihy	9
1.2	Komu je kniha určena	11
1.3	Způsob psaní a použité konvence	11
1.4	Práce s knihou	11
I	Východiska a související témata	15
2	Dobývání znalostí z databází	17
2.1	Vznik a cíle	17
2.2	Základní fáze DZD	20
2.3	Porozumění doménové oblasti	23
2.4	Porozumění datům	25
2.5	Předzpracování dat	27
2.6	Modelování a analytické procedury	35
2.7	Vyhodnocení výsledků	38
2.8	Využití výsledků	39
3	Relevantní témata DZD	41
3.1	Data a matice dat	41
3.2	Nominální, ordinální a kardinální atributy	43
3.3	Asociační pravidla	43
3.4	Přehled ostatních relevantních témat	44
4	Metoda GUHA a logika objevování pomocí počítače	45
4.1	GUHA-procedury	46
4.2	Booleovské atributy	46
4.3	GUHA-procedura ASSOC	48
4.4	Logika objevování pomocí počítače	52
4.5	Observační kalkul asociačních pravidel	53
4.6	Třídy 4ft-kvantifikátorů	54
4.7	Dedukční pravidla v kalkulu asociačních pravidel	56
4.8	Neúplná informace	58
4.9	Implementace metody GUHA	62

5	Projekt LISp-Miner	65
5.1	Historie vývoje systému	65
5.2	Projektové řízení	71
5.3	Využití systému LISp-Miner ve výuce a v praxi	73
 II Metoda GUHA a systém LISp-Miner		75
6	Metoda GUHA a systém LISp-Miner	77
6.1	Přehled GUHA-procedur systému LISp-Miner	78
6.2	Společné prvky zadání procedur	80
6.3	Terminologie, vztah k observačním kalkulům	81
6.4	Podmíněné vztahy a kontingenční tabulky	82
6.5	Kvantifikátory	83
6.6	Množina relevantních booleovských atributů	85
7	GUHA-procedura 4ft-Miner	89
7.1	Přehled	89
7.2	Čtyřpolní tabulky a podmíněné čtyřpolní tabulky	90
7.3	Statistické 4ft-kvantifikátory	91
7.4	Jednoduché frekvenční 4ft-kvantifikátory	93
7.5	4ft-kvantifikátory BASE a Ceiling	94
7.6	Podmíněná asociační pravidla	95
7.7	Podmíněná pravidla a neúplná informace	97
7.8	Parametry procedury	98
8	GUHA-procedura CF-Miner	103
8.1	Přehled	103
8.2	CF-tabulka	104
8.3	Distribuční CF-kvantifikátory	105
8.4	Výsek CF-tabulky	106
8.5	Jednoduché frekvenční CF-kvantifikátory	107
8.6	Existenční CF-kvantifikátory	109
8.7	Schody v histogramu	110
8.8	CF-kvantifikátory pro schody v histogramu	112
8.9	CF-Miner a neúplná informace	113
9	GUHA-procedura KL-Miner	115
9.1	Přehled	115
9.2	KL-tabulky a podmíněné KL-tabulky	116
9.3	Výsek KL-tabulky	117
9.4	Pokročilejší KL-kvantifikátory	118
9.5	Jednoduché frekvenční KL-kvantifikátory	119
9.6	Existenční KL-kvantifikátory	120
9.7	KL-Miner a neúplná informace	121
10	GUHA-procedura MCluster-Miner	123
10.1	Shluková analýza	124

10.2	Podobnost dvou objektů	125
10.3	Míry podobnosti	125
10.4	Zobrazení vícerozměrných dat	127
10.5	Identifikace shluků	128
10.6	Posouzení kvality shlukování	128
10.7	Způsoby shlukování	129
10.8	Shlukování na diskretizovaných datech	131
10.9	Podmíněná shluková analýza	132
11	GUHA-procedura ETree-Miner	133
11.1	Rozhodovací stromy	134
11.2	Strom jako model analyzovaných dat	137
11.3	Konstrukce rozhodovacího stromu	137
11.4	Podmíněné rozhodovací stromy	139
11.5	Explorační stromy	139
11.6	Testování kvality stromu	141
12	GUHA-procedury typu SD porovnávající podmnožiny záznamů	143
12.1	Procedury pro dvojice podmnožin	143
12.2	GUHA-procedura SD4ft-Miner	144
12.3	GUHA-procedura SDCF-Miner	149
12.4	GUHA-procedura SDKL-Miner	156
12.5	GUHA-procedura Ac4ft-Miner	162
III	Systém LISp-Miner a proces DZD	169
13	Základy práce se systémem LISp-Miner	171
13.1	Instalace systému	171
13.2	Práce se systémem	173
13.3	Modul LM Workspace	174
14	Seznámení s daty v systému LISp-Miner	179
14.1	Načtení dat do systému LISp-Miner	179
14.2	Prohlížení dat	183
14.3	Slučování sloupců rozsáhlých tabulek	188
15	Předzpracování dat v systému LISp-Miner	191
15.1	Strom skupin a do nich patřící atributy	192
15.2	Kategoriální atribut	194
15.3	Automatické vytváření kategorií	196
15.4	Ruční editace kategorií	200
15.5	Vytvoření nového atributu	201
15.6	Klonování atributu	201
15.7	Vhodné pojmenování atributů a kategorií	202
15.8	Předzpracování dat s neúplnou informací	203
15.9	Dichotomizace atributu	204
15.10	Datová matice	206

15.11	Propojení dat z více tabulek do jedné	206
15.12	Výpočet odvozených hodnot	208
15.13	Sloupce založené na výsledcích úloh	209
16	Interaktivní analýza v systému LISp-Miner	211
16.1	Scatter-plot analýza	211
16.2	Analýza hlavních komponent	213
16.3	Frekvenční analýza kategorií	214
16.4	Korelační analýza atributů	216
16.5	Kontingenční analýza jednoho atributu	216
16.6	Kontingenční analýza dvou atributů	219
16.7	Kontingenční analýza booleovských atributů	221
16.8	Interaktivní vytváření rozhodovacího stromu	222
17	Analytické modelování v systému LISp-Miner	225
17.1	Seznam úloh	226
17.2	Detail zadání úlohy	227
17.3	Kontrola zadání úlohy	229
17.4	Vytvoření a klonování úlohy	230
17.5	Výpočet úlohy	231
17.6	Prohlížení výsledků úlohy	234
17.7	Souhrnné výsledky všech úloh	236
18	Zadávání atributů a množiny relevantních cedentů	237
18.1	Výběr atributů	237
18.2	Zadání množiny relevantních cedentů	239
18.3	Zadání množiny relevantních dílčích cedentů	241
18.4	Zadání množiny relevantních literálů	243
18.5	Kontrola zadání cedentu	244
18.6	Tipy a triky pro zadávání cedentů	245
18.7	Zadání prostého seznamu atributů	246
IV	Analytické moduly systému LISp-Miner	247
19	Modul 4ft-Miner	249
19.1	Vytvoření úlohy	249
19.2	Příklad aplikace	255
19.3	Kontrola zadání úlohy	260
19.4	Přidání odvozeného sloupce nad pravidlem	260
20	Modul CF-Miner	261
20.1	Vytvoření úlohy	261
20.2	Příklad aplikace procedury CF-Miner	269
20.3	Kontrola zadání úlohy	272
20.4	Přidání odvozeného sloupce nad CF-vztahem	272

21 Modul KL-Miner	273
21.1 Vytvoření úlohy	273
21.2 Příklad aplikace procedury KL-Miner	279
21.3 Kontrola zadání úlohy	281
21.4 Přidání odvozeného sloupce nad KL-vztahem	282
22 Modul MCluster-Miner	283
22.1 Vytvoření úlohy pro proceduru MCluster-Miner	283
22.2 Výpočet úlohy	286
22.3 Prohlížení výsledků	288
22.4 Detail varianty shlukování	289
22.5 Příklad úlohy pro MCluster-Miner	291
22.6 Tipy a triky pro MCluster-Miner	294
23 Modul ETree-Miner	295
23.1 Vytvoření úlohy pro proceduru ETree-Miner	295
23.2 Výpočet úlohy	298
23.3 Prohlížení výsledků	300
23.4 Detail stromu	301
23.5 Příklad úlohy pro ETree-Miner	303
23.6 Tipy a triky pro ETree-Miner	306
24 Moduly typu SD porovnávající podmnožiny záznamů	307
24.1 Modul SD4ft-Miner	307
24.2 Modul SDCF-Miner	312
24.3 Modul SDKL-Miner	317
24.4 Modul Ac4ft-Miner	321
V Systém LISp-Miner – pokročilá témata	325
25 Systém LISp-Miner	327
25.1 Požadavky a výchozí předpoklady	327
25.2 Celková koncepce	328
25.3 Návrh systému	331
25.4 Implementace systému	332
26 Skriptovací jazyk LMCL a modul LM Exec	335
26.1 Jazyk Lua	336
26.2 LISp-Miner Control Language	337
26.3 Modul LM Exec	340
26.4 Dokumentace LMCL	341
26.5 Ukázkový příklad EverMinerSimple Demo	342
27 Generování a verifikace asociačních pravidel	347
27.1 Generování hypotéz v proceduře 4ft-Miner	347
27.2 Verifikace relevantních otázek	355
27.3 Rychlost generování a verifikace	358

28 Bitové řetězce	359
28.1 Bitové řetězce jako datová struktura	359
28.2 Operace nad bitovými řetězci	361
28.3 Bitové řetězce a nároky na operační paměť	363
28.4 Bitové řetězce a rychlost výpočtu	363
29 Dávkový výpočet úloh	367
29.1 Modul LM TaskPooler	367
29.2 Modul LM ProcPooler	370
30 Distribuovaný výpočet úloh na počítačovém gridu	373
30.1 Dosavadní algoritmus generování a verifikace	373
30.2 Možné strategie distribuovaného výpočtu	374
30.3 Kroky distribuovaného výpočtu na gridu	379
30.4 Implementace a nasazení	381
30.5 Modul LM GridPooler	384
31 Výměna dat s jinými systémy	387
31.1 Export dat	387
31.2 Import dat	392
32 Generování umělých dat a modul LM ReverseMiner	395
32.1 Východiska generování umělých dat	396
32.2 Existující přístupy ke generování umělých dat	396
32.3 Navržený přístup	397
32.4 Evoluční operace	399
32.5 Modul ReverseMiner	402
32.6 Vytvoření dat pro tuto publikaci	404
32.7 Přednosti, omezení a další vývoj	406
33 Výzkumné projekty	407
33.1 Projekt EverMiner	408
33.2 Projekt FOFRADOC – popis procesu DZD	410
33.3 Dílčí automatizace procesu DZD	422
VI Přílohy	427
34 Popis použitých datových sad	429
34.1 Data STULONG	429
34.2 Data Iris	430
Použitá literatura a zdroje	431
Rejstřík	451