

Tartalomjegyzék

Előszó a magyar kiadáshoz	13
Előszó az angol kiadáshoz	15
Szerzők előszava	17
1. fejezet Bevezetés	23
1.1. Az adatbányászat kialakulása és fontossága	23
1.2. Mi az adatbányászat?	26
1.3. Adatbányászat – az adatok típusa	31
1.3.1. <i>Relációs adatbázisok</i>	31
1.3.2. <i>Adattárházak</i>	33
1.3.3. <i>Tranzakciós adatbázisok</i>	36
1.3.4. <i>Fejlett adatbázisrendszerek és adatbázis-alkalmazások</i>	37
1.4. Milyen minták bányászhatók?	42
1.4.1. <i>Fogalom-/osztályleírások: karakterizáció és diszkrimináció</i>	42
1.4.2. <i>Társításelemzés</i>	44
1.4.3. <i>Osztályozás és előrejelzés</i>	45
1.4.4. <i>Klaszterelemzés</i>	46
1.4.5. <i>Szélsőséges értékek elemzése</i>	47
1.4.6. <i>Fejlődésanalízis</i>	47
1.5. Minden minta érdekes?	48
1.6. Az adatbányászati rendszerek osztályozása	49
1.7. Az adatbányászat fő kérdései	51
1.8. Összefoglalás	54
1.9. Feladatok	55
1.10. Irodalom	56
2. fejezet Adattárház és OLAP-technológia	59
2.1. Mi az adattárház?	59
2.1.1. <i>Különbségek az operatív adatbázisrendszerek és az adattárházak között</i>	62
2.1.2. <i>Miért legyen mégis külön adattárház?</i>	63

2.2.	A többdimenziós adatmodell	64
2.2.1.	<i>A táblázattól és számológéptől az adatkockáig</i>	64
2.2.2.	<i>Csillagok, hópehelyek és csillagképek: sémák többdimenziós adatbázisokhoz</i>	68
2.2.3.	<i>Példák csillag-, hópehely- és csillagképsémák definiálására</i>	70
2.2.4.	<i>A mértékek osztályozása és számítása</i>	73
2.2.5.	<i>Fogalmi hierarchiák bevezetése</i>	75
2.2.6.	<i>OLAP-műveletek a többdimenziós adatmodellben</i>	77
2.2.7.	<i>Csillaghálós lekérdező modell többdimenziós adatbázisok lekérdezéséhez</i>	80
2.3.	Adattárház-architektúra	81
2.3.1.	<i>Adattárház tervezésének és építésének lépései</i>	81
2.3.2.	<i>A háromszintű adattárház felépítése</i>	84
2.3.3.	<i>Az OLAP-szerverek típusai: ROLAP, MOLAP, HOLAP</i>	87
2.4.	Adattárházak megvalósítása	89
2.4.1.	<i>Adatkockák hatékony számítása</i>	89
2.4.2.	<i>OLAP-adatok indexelése</i>	96
2.4.3.	<i>OLAP-lekérdezések hatékony feldolgozása</i>	99
2.4.4.	<i>Metaadatraktár</i>	100
2.4.5.	<i>Az adattárházak háttéreszközei és segédprogramjai</i>	101
2.5.	Az adatkocka-technológia továbbfejlesztése	102
2.5.1.	<i>Az adatkockák felfedezésvezérelt feltárása</i>	102
2.5.2.	<i>Bonyolult összesítések több finomsági szinten: többtulajdonságú kockák</i>	106
2.5.3.	<i>További fejlesztések</i>	109
2.6.	Az adattárház-kezeléstől az adatbányászatig	109
2.6.1.	<i>Adattárházak használata</i>	109
2.6.2.	<i>Az on-line analitikus feldolgozástól az on-line analitikus bányászatig</i>	111
2.7.	Összefoglalás	114
2.8.	Feladatok	115
2.9.	Irodalom	118
3.	fejezet Az adatok előfeldolgozása	121
3.1.	Az adatok előfeldolgozásának szükségessége	121
3.2.	Adattisztítás	125
3.2.1.	<i>Hiányzó értékek</i>	125
3.2.2.	<i>Zajos adatok</i>	126
3.2.3.	<i>Inkonzisztens adatok</i>	128
3.3.	Adatok integrálása és transzformálása	128
3.3.1.	<i>Adatok integrálása</i>	128
3.3.2.	<i>Adatok transzformálása</i>	130
3.4.	Adatok redukálása	132
3.4.1.	<i>Összevonás adatkockába</i>	133
3.4.2.	<i>Dimenziócsökkentés</i>	134

3.4.3.	<i>Adatok tömörítése</i>	137
3.4.4.	<i>Számosságcsökkentés</i>	140
3.5.	Diszkretizáció és fogalmi hierarchiák generálása	146
3.5.1.	<i>Diszkretizáció és fogalmi hierarchiák generálása numerikus adatokból</i>	147
3.5.2.	<i>Fogalmi hierarchiák generálása kategória típusú adatokból</i>	152
3.6.	Összefoglalás	155
3.7.	Feladatok	155
3.8.	Irodalom	156

4. fejezet | Adatbányászó primitívek, nyelvek és rendszerarchitektúrák 159

4.1.	Adatbányászó primitívek: mit jelent az adatbányászati feladat?	160
4.1.1.	<i>A feladat szempontjából fontos adatok</i>	162
4.1.2.	<i>A bányászandó tudás típusa</i>	163
4.1.3.	<i>Háttértudás: fogalmi hierarchiák</i>	164
4.1.4.	<i>Érdekességi mértékek</i>	167
4.1.5.	<i>A felfedezett mintázat bemutatása és ábrázolása</i>	170
4.2.	Az adatbányászati lekérdező nyelv	172
4.2.1.	<i>A feladat szempontjából fontos adatok meghatározásának szintaktikája</i>	174
4.2.2.	<i>Szintaktika a bányászandó tudás típusának meghatározására</i>	175
4.2.3.	<i>A fogalmi hierarchia meghatározásának szintaktikája</i>	177
4.2.4.	<i>Az érdekességi mérték meghatározásának szintaktikája</i>	179
4.2.5.	<i>Szintaxis a mintázat bemutatásának és ábrázolásának meghatározására</i>	179
4.2.6.	<i>Mindent egybevéve – példa a DMQL-lekérdezésre</i>	180
4.2.7.	<i>További adatbányászati nyelvek és az adatbányászó primitívek szabványosítása</i>	182
4.3.	Az adatbányászati lekérdező nyelven alapuló grafikus felhasználói interfész fejlesztése	183
4.4.	Az adatbányászati rendszerek architektúrája	184
4.5.	Összefoglalás	186
4.6.	Feladatok	187
4.7.	Irodalom	189

5. fejezet | Fogalomleírás: jellemzés és összehasonlítás 191

5.1.	A fogalomleírás	191
5.2.	Adatáltalánosítás és összegzés alapú jellemzés	193
5.2.1.	<i>Az attribútumorientált indukció</i>	193
5.2.2.	<i>Az attribútumorientált indukció hatékony megvalósítása</i>	199
5.2.3.	<i>A kapott általánosítás megjelenítése</i>	201
5.3.	Analitikus jellemzés: attribútumrelevancia-elemzés	205
5.3.1.	<i>Miért végzünk attribútumrelevancia-elemzést?</i>	205
5.3.2.	<i>Az attribútumrelevancia elemzésének módszerei</i>	207
5.3.3.	<i>Analitikus jellemzés – példa</i>	209

5.4.	Osztály-összehasonlítások bányászata: különböző osztályok megkülönböztetése	211
5.4.1.	<i>Osztály-összehasonlítási módszerek és megvalósításuk</i>	211
5.4.2.	<i>Az osztály-összehasonlítási leírások megjelenítése</i>	214
5.4.3.	<i>Osztályleírás: jellemzés és összehasonlítás együttes megjelenítése</i>	216
5.5.	Leíró statisztikai mértékek bányászata nagy adatbázisokban	218
5.5.1.	<i>Az elhelyezkedés mérése</i>	218
5.5.2.	<i>Az adatszóródás mérése</i>	220
5.5.3.	<i>Az alapvető statisztikai osztályleírások grafikus megjelenítése</i>	222
5.6.	Tárgyalás	227
5.6.1.	<i>Fogalomleírás: tipikus gépi tanulási módszerekkel történő összehasonlítás</i>	227
5.6.2.	<i>A fogalomleírás növekményes és párhuzamos bányászata</i>	229
5.7.	Összefoglalás	230
5.8.	Feladatok	231
5.9.	Irodalom	232

6. fejezet | Társítási szabályok bányászata nagy adatbázisokban 233

6.1.	Társítási szabályok bányászata	233
6.1.1.	<i>Vásárlói kosár elemzése – példa, amely motiválta a társítási szabályok bányázatát</i>	234
6.1.2.	<i>Alapfogalmak</i>	235
6.1.3.	<i>A társítási szabályok bányázatának feltérképezése</i>	236
6.2.	Az egydimenziós Boole társítási szabályok bányászata tranzakciós adatbázisokban	238
6.2.1.	<i>Az Apriori algoritmus: gyakori elemhalmazok keresése jelöltek előállításával</i>	238
6.2.2.	<i>Társítási szabályok generálása a gyakori elemhalmazokból</i>	243
6.2.3.	<i>Az Apriori algoritmus hatékonyságának növelése</i>	244
6.2.4.	<i>Gyakori elemhalmazok keresése jelöltek generálása nélkül</i>	246
6.2.5.	<i>Jéghegy típusú kérdések</i>	250
6.3.	Többszintű társítási szabályok bányászata tranzakciós adatbázisokban	251
6.3.1.	<i>Többszintű társítási szabályok</i>	251
6.3.2.	<i>A többszintű társítási szabályok bányázatáról</i>	253
6.3.3.	<i>Többszintű társítási szabályok feleslegességének ellenőrzése</i>	257
6.4.	Többdimenziós társítási szabályok bányászata relációs adatbázisokban és adattárházakban	258
6.4.1.	<i>Többdimenziós társítási szabályok</i>	258
6.4.2.	<i>Többdimenziós társítási szabályok bányászata a mennyiségi attribútumok statikus diszkrétizációjával</i>	260
6.4.3.	<i>A mennyiségi társítási szabályok bányászata</i>	261
6.4.4.	<i>A társítási szabályok távolság alapú bányászata</i>	264
6.5.	A társítás bányászásától a korrelációs analízisig	266
6.5.1.	<i>Az erős szabályok nem feltétlenül érdekesek – példa</i>	266
6.5.2.	<i>A társításelemzéstől a korrelációs analízisig</i>	267

6.6.	Társítások bányászata megszorításokkal	269
6.6.1.	<i>Társítási szabályok metasabály-vezérelt bányászata</i>	269
6.6.2.	<i>Kiegészítő szabálymegszorítások által vezérelt bányászás</i>	271
6.7.	Összefoglalás	275
6.8.	Feladatok	277
6.9.	Irodalom	282

7. fejezet | Osztályozás és előrejelzés 285

7.1.	Az osztályozás és az előrejelzés	285
7.2.	Témakörök az osztályozásra és előrejelzésre vonatkozóan	288
7.2.1.	<i>Az adatok előkészítése osztályozásra, illetve előrejelzésre</i>	288
7.2.2.	<i>Az osztályozási módszerek összehasonlítása</i>	289
7.3.	Osztályozás a döntési fa indukció segítségével	289
7.3.1.	<i>Döntési fa indukció</i>	290
7.3.2.	<i>Fametszés</i>	295
7.3.3.	<i>Az osztályozási szabályok kinyerése döntési fákból</i>	296
7.3.4.	<i>A döntési fa indukció alapmódszerének javítása</i>	297
7.3.5.	<i>Skálázhatóság és a döntési fa indukció</i>	298
7.3.6.	<i>Az adattárházak technikák és a döntési fa indukció integrálása</i>	300
7.4.	Bayes-osztályozás	301
7.4.1.	<i>Bayes-tétel</i>	302
7.4.2.	<i>Naiv Bayes-osztályozó</i>	303
7.4.3.	<i>Bayes-féle hihetőségi hálók</i>	305
7.4.4.	<i>A Bayes-féle hihetőségi hálók tanítása</i>	306
7.5.	Osztályozás hiba-visszaterjesztéssel	308
7.5.1.	<i>Előrecsatolt többrétegű hálózatok</i>	308
7.5.2.	<i>A háló topológiájának definiálása</i>	309
7.5.3.	<i>A hiba-visszaterjesztés algoritmus</i>	310
7.5.4.	<i>A hiba-visszaterjesztés algoritmus és az értelmezhetőség</i>	315
7.6.	A társítási szabályok bányászatán alapuló osztályozás	316
7.7.	Más osztályozási módszerek	318
7.7.1.	<i>A k-legközelebbi szomszédon alapuló osztályozás</i>	318
7.7.2.	<i>Eset alapú következtetés</i>	319
7.7.3.	<i>Genetikus algoritmusok</i>	320
7.7.4.	<i>Közelítő halmazokon alapuló megközelítés</i>	321
7.7.5.	<i>Fuzzy halmazos megközelítések</i>	322
7.8.	Előrejelzés	323
7.8.1.	<i>Lineáris és többváltozós regresszió</i>	323
7.8.2.	<i>Nemlineáris regresszió</i>	325
7.8.3.	<i>Más regressziós modellek</i>	326
7.9.	Az osztályozó pontossága	327
7.9.1.	<i>Az osztályozó pontosságának becslése</i>	327
7.9.2.	<i>Az osztályozó pontosságának növelése</i>	328
7.9.3.	<i>Elegendő csupán a pontosság ismerete egy-egy osztályozó megítéléséhez?</i>	329

7.10. Összefoglalás	331
7.11. Feladatok	332
7.12. Irodalom	334
8. fejezet Klaszterelemzés	339
8.1. A klaszterelemzés	339
8.2. Adattípusok a klaszterelemzés során	342
8.2.1. <i>Intervallumváltozók</i>	343
8.2.2. <i>Bináris változók</i>	345
8.2.3. <i>Felsorolás típusú, rendezett arányskálázott változók</i>	347
8.2.4. <i>Vegyes típusú változók</i>	349
8.3. A legfontosabb klaszterező módszerek osztályozása	350
8.4. Particionáló módszerek	353
8.4.1. <i>Klasszikus particionáló módszerek: k-átlag és k-medoid</i>	353
8.4.2. <i>Particionáló módszerek nagy adatbázisokban: a k-medoidtól a CLARANS-ig</i>	357
8.5. Hierarchikus módszerek	358
8.5.1. <i>Egyesítő és felosztó hierarchikus klaszterezés</i>	359
8.5.2. <i>BIRCH: kiegyensúlyozott iteratív csökkentés és klaszterezés hierarchiák segítségével</i>	360
8.5.3. <i>CURE: klaszterezés reprezentáló elemek segítségével</i>	362
8.5.4. <i>Chameleon: dinamikus modellezést használó hierarchikus klaszterező algoritmus</i>	364
8.6. Sűrűség alapú módszerek	366
8.6.1. <i>DBSCAN: megfelelően sűrű és összefüggő területekre alapozó, sűrűség alapú klaszterezési módszer</i>	366
8.6.2. <i>OPTICS: a pontok rendezése a klaszterező struktúra azonosításához</i>	368
8.6.3. <i>DENCLUE: klaszterezés sűrűségeloszlás-függvények alapján</i>	370
8.7. Rács alapú módszerek	372
8.7.1. <i>STING: statisztikai információs rács</i>	373
8.7.2. <i>WaveCluster: klaszterezés hullámtranszformáció segítségével</i>	374
8.7.3. <i>CLIQUE: sokdimenziós terek klaszterezése</i>	376
8.8. Modell alapú klaszterező módszerek	378
8.8.1. <i>Statisztikai megközelítés</i>	378
8.8.2. <i>Neuronháló megközelítés</i>	381
8.9. Szélsőséges értékek elemzése	383
8.9.1. <i>Szélsőséges értékek statisztikai alapú keresése</i>	384
8.9.2. <i>Távolság alapú szélsőséges értékek keresése</i>	386
8.9.3. <i>Szélsőséges értékek eltérés alapú keresése</i>	387
8.10. Összefoglalás	390
8.11. Feladatok	391
8.12. Irodalom	393

9. fejezet Komplex adattípusok bányászata	395
9.1. Komplex adatobjektumok többdimenziós analízise és leíró bányászata	395
9.1.1. <i>Strukturált adatok általánosítása</i>	396
9.1.2. <i>Összesítés és approximáció a térbeli és a multimédia-adatok általánosításában</i>	397
9.1.3. <i>Az objektumazonosítók és az osztály-álosztály hierarchiák általánosítása</i>	398
9.1.4. <i>Az osztályszerkezet-hierarchiák általánosítása</i>	399
9.1.5. <i>Objektumkockák létrehozása és bányászata</i>	399
9.1.6. <i>Tervadatbázisok általánosítás alapú bányászata az „oszd meg és uralkodj” módszerrel</i>	400
9.2. Téradatbázisok bányászata	404
9.2.1. <i>Téradatkocka létrehozása és tér-OLAP</i>	404
9.2.2. <i>Térbeli társításelemzés</i>	409
9.2.3. <i>Térbeli klaszterezési módszerek</i>	410
9.2.4. <i>Térbeli osztályozás és térbeli trendanalízis</i>	410
9.2.5. <i>Raszteres adatbázisok bányászata</i>	411
9.3. Multimédia-adatbázisok bányászata	411
9.3.1. <i>Hasonlóság keresése a multimédia-adatokban</i>	411
9.3.2. <i>Multimédia-adatok többdimenziós analízise</i>	413
9.3.3. <i>Multimédia-adatok osztályozása és előrejelzés-analízise</i>	415
9.3.4. <i>Társításbányászat a multimédia-adatokban</i>	415
9.4. Idősorok és szekvenciális adatok bányászata	416
9.4.1. <i>Trendanalízis</i>	417
9.4.2. <i>Hasonlóság keresése az idősorok elemzésében</i>	419
9.4.3. <i>Szekvenciális minták bányászata</i>	423
9.4.4. <i>Ismétlődésanalízis</i>	424
9.5. Szöveges adatbázisok bányászata	426
9.5.1. <i>Elemzés és információvisszakeresés szöveges adatokon</i>	426
9.5.2. <i>Szövegbányászat: kulcsszó alapú társítás és dokumentumosztályozás</i>	431
9.6. A Word Wide Web bányászata	433
9.6.1. <i>A weblink struktúráinak bányászata az autentikus weblapok lokalizálására</i>	435
9.6.2. <i>Webdokumentumok automatikus osztályozása</i>	437
9.6.3. <i>Többretegű webinformációs bázis létrehozása</i>	438
9.6.4. <i>Webhasználat bányászata</i>	439
9.7. Összefoglalás	440
9.8. Feladatok	441
9.9. Irodalom	444

10. fejezet Alkalmazások és irányzatok az adatbányászatban	447
10.1. Az adatbányászat felhasználási területei	447
10.1.1. <i>Orvostudományi és DNS-adatok bányászata</i>	447
10.1.2. <i>Adatbányászat a pénzügyekben</i>	449
10.1.3. <i>Adatbányászat a kiskereskedelemben</i>	451
10.1.4. <i>Adatbányászat a távközlésben</i>	452
10.2. Adatbányászati termékek és kutatási fázisban lévő prototípusok	453
10.2.1. <i>Az adatbányászati rendszer megválasztása</i>	454
10.2.2. <i>Néhány adatbányászati termék bemutatása</i>	456
10.3. További adatbányászati témák	458
10.3.1. <i>Látásra és hallásra támaszkodó adatbányászat</i>	458
10.3.2. <i>Tudományos és statisztikai adatbányászat</i>	463
10.3.3. <i>Az adatbányászat elméleti alapjai</i>	464
10.3.4. <i>Adatbányászat és intelligens válaszadás</i>	466
10.4. Az adatbányászat társadalmi hatásai	467
10.4.1. Felfújtt léggömb vagy folyamatosan és biztosan növekedő üzletág?	467
10.4.2. <i>Adatbányászat: menedzsereknek vagy mindenkinek?</i>	469
10.4.3. <i>Fenyegeti-e az adatbányászat az adatbiztonságot és személyes titkainkat?</i>	470
10.5. Trendek az adatbányászatban	473
10.6. Összefoglalás	474
10.7. Feladatok	475
10.8. Irodalom	477
A) függelék A Microsoft OLE DB for Data Mining szabványa	479
A.1. DMM-objektum létrehozása	480
A.2. Betanítási adatok beszúrása a modellbe és a modell betanítása	481
A.3. A modell használata	482
B) függelék Bevezetés a DBMinerbe	487
B.1. A rendszer felépítése	488
B.2. Input – output	489
B.3. A rendszer által támogatott adatbányászati feladatok	489
B.4. Feladat- és módszerválasztás	492
B.5. KDD-folyamat	492
B.6. Fontosabb alkalmazások	493
B.7. A jelenlegi helyzet	493
Irodalomjegyzék	495
Tárgymutató	517