

INFORMATIKAI KAR DOKTORI PROGRAM
SZIGORLATI TEMATIKA

Adatbányászat (melléktárgy)

Az adatbányászat feladata: A tudásfeltárás és az adatbányászat felhasználási területei. Adattárházak, OLAP-adatok, OLAP-műveletek a többdimenziós adatmodellben. Az adatok előfeldolgozása, az adatok tisztítása, az adatok integrálása, transzformálása, redukálása, tömörítése. Diszkrétizáció, fogalmi hierarchiák. Adatbányászati modellek általános jellemzői, társítási (asszociációs) szabályok, klaszterezés, döntési fák, osztályozás, előrejelzés, neuronhálók, regresszió.

Társítási (asszociációs) szabályok: Vásárlói kosarak elemzése, gyakori halmazok, gyakori minták, az Apriori algoritmus, hatékonyság növelése, a PCY (Park, Chen, Yu) algoritmus, Toivonen algoritmus, DIC algoritmus, Zaki algoritmus, FUP algoritmus, Apriori TID algoritmus, DHP algoritmus, FP algoritmus, jéghegy típusú kérdések. Töredelő (hash) technikák, MinHashing, Locality-Sensitive Hashing, k-minHashing, Hamming LSH. Többdimenziós társítási szabályok, korrelációs analízis.

Szekvenciabányászat és epizódbányászat: Szekvenciaillesztés, gyakori sorozatok, indexelés Fourier transzformáció segítségével, időkénszerek, GSP algoritmus. Gyakori epizódok, monotonitás, az Apriori algoritmus alkalmazása, párhuzamos és soros epizódok ellenőrzése.

Osztályozás és előrejelzés: Döntési fák, ID3 algoritmus, C4.5 algoritmus, Bayes-osztályozás, Bayes-hálók, neuronhálók, hiba-visszaterjesztés, k-legközelebbi szomszédon alapuló osztályozás, genetikus algoritmusok, lineáris és nemlineáris regresszió, többdimenziós regresszió.

Klaszterelemzés: A klaszterek jellemzői, távolság, átmérő, k-közép, legkisebb négyzetes hiba. Klaszterező módszerek, particionáló, hierarchikus, spektrál-alapú, sűrűség-alapú, grid-alapú módszerek, k-átlag algoritmus, BFR algoritmus, dimenziócsökkentés, PAM, CLARA, CLARANS algoritmusok. AGNES, DIANA, BIRCH, CURE, Chameleon, DBSCAN, GRGPF, OPTICS algoritmusok.

Webes adatbányászat: Az oldalak rangja, webnapló, zsácutca és pókháló probléma, Google módszer, SALSA módszer, gyűjtőlapok, tekintélyek. Szokatlan minták, DICE módszer.

Adatbányászat a gyakorlatban: Adatbányászati szabványok, irányzatok, DMM, adatbányászati primitívek, adatbányászati lekérdező nyelvek, grafikus felületek. Adatbányászati szoftverek, DBMiner, SAS Enterprise Miner, IBM Intelligent Miner, ISL Clementine, SGI MineSet. Adattípusok, rendszerkövetelmények, adatforrások, adatbányászati funkciók.

Irodalom:

J. Han, M. Kamber: Adatbányászat, PANEM, Budapest, 2004.

Bodon Ferenc, Lukács András: Adatbányászat, Informatikai algoritmusok. 2. kötet, 33. fejezet, ELTE Eötvös Kiadó, Budapest, 2005.

Bodon Ferenc: Adatbányászati algoritmusok, (tanulmány), Budapest, 2004.

<http://www.cs.bme.hu/~bodon/magyar/adatbanyaszat/tanulmany/adatbanyaszat.pdf>

J.D.Ullman: Datamining, CS345 Lecture Notes, Stanford University, USA, 2000.

(<http://www-db.stanford.edu/~ullman/mining/allnotes.pdf>)

G. Chang, M.J.Healey, J.A.M. McHugh, J.T.L. Wang: Mining the World Wide Web, Kluwer Academic Publisher, 2001.

Usama M. Fayyadm G.P.Shapiro, P. Smyth, R. Uthurusamy: Advances in Knowledge Discovery and Data Mining, AAAI Press/ The MIT Press, 1996.

P. Adriaans, D. Zantinge: Adatbányászat, PANEM, Budapest, 2002.
