Využitie procesu získavania znalostí z databáz pri riadení procesov

Using of Knowledge discovery data in control process

Andrej Trnka, Katedra aplikovanej informatiky FPV UCM v Trnave

Abstract: Paper describes possibilities of using datawarehouses and method of Knowledge discovery data in control proces.

Key words: datawarehouse, control proces, knowledge discovery data, data mining

Abstrakt: Článok popisuje možnosti využitia dátových skladov a techník dobývania znalostí z databáz v riadiacom procese..

Kľúčové slová: dátový sklad, riadiaci proces, dobývanie znalostí z databáz, dolovanie dát

1. Úvod

Riadenie procesov je oblasť automatizácie, pri ktorej sa môžeme stretnúť s veľmi veľkým množstvom údajov. Tieto údaje môžu, ale aj nemusia byť pre riadiaci proces relevantné. Úlohou využitia dátových skladov pri riadení procesov je tieto dáta uchovávať v tzv. dimenziách (multidimenzionálny prístup). Výhodou multidimenzionálneho prístupu (OLAP) oproti relačnému (OLTP) je to, že údaje z riadiaceho procesu je možné uchovať nie v dvojrozmerných tabuľkách, ale je ich možné uchovať v tzv. OLAP kocke, v ktorej je možné sa pohybovať vo viacerých dimenziách (napr. čas, miesto, veličina). Keďže v dátovom sklade sú uložené rôznorodé údaje, je potrebné z týchto údajov získať potrebné informácie, ktoré je potom možné použiť na zlepšenie riadiaceho procesu.

2. Dátový sklad

Dátový sklad je subjektovo orientovaná, integrovaná, stála a časovo rozlíšená kolekcia dát určená na podporu procesu rozhodovania. V dátovom sklade sú uložené atomické a sumárne dáta. Údaje sa získavajú a ukladajú do produkčných (operačných) databáz, ktoré môžu byť na rozličných lokalitách. Tieto údaje sa v pravidelných intervaloch zozbierajú, predspracujú a zavedú do dátového skladu (ETL). Dátový sklad je v podstate takisto databáza, len je

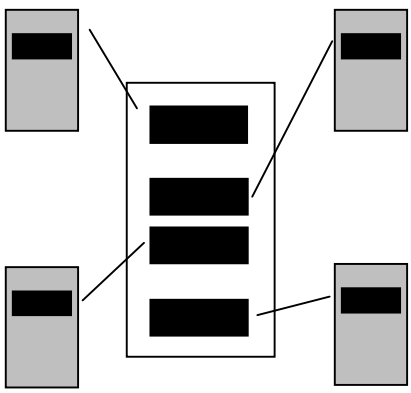
organizovaná podľa iných pravidiel ako transakčná databáza. Napr. tabuľky nemusia byť normalizované a majú multidimenzionálnu štruktúru. Dátový sklad je teda súbor technológií na efektívne skladovanie údajov tak, aby tieto údaje po ich premene na informácie slúžili na podporu rozhodovania. (1)

Dátový sklad obsahuje:

- Jasné a integrované dáta - umožnenie jednoduchého pohľadu na dáta bez rušivých záležitostí (normalizácia, kľúčové determinanty, chýbajúce hodnoty...).
- Detailné a sumárne dáta - detailné dáta sú potrebné pri pohľade na dáta, ktoré sú vo forme najväčšej granularity a je potrené extrahovať dôležité vzorky. Sumárne dáta sú dôležité pre učenie sa o vzorkách v dátach, ktoré už boli extrahované niekým iným. Tieto dáta zaisťujú to, že pri dolovaní údajov nie je potrené začínať od začiatku, ale je možné pokračovať tam, kde niekto iný skončil.
- Historické dáta - tieto dáta pomáhajú analyzovať trendy v minulosti.
- Metadáta - sú používané k popisu súvislostí a významu dát. (2)

2.1. Multidimenzionálny model

Multidimenzionálny prístup je niekedy nazývaný tiež hviezdicový. Stredobodom multidimenzionálneho prístupu k návrhu databáze je hviezdicové prepojenie, ktoré je zobrazené na obr. 1.



Obrázok 1: Hviezdicové prepojenie

Štruktúra dát sa volá hviezdicová preto, lebo jej reprezentácia zobrazuje hviezdu so stredom a niekoľkými odľahlými štruktúrami dát.

Pri multidemnzionálnom prístupe sa hovorí o faktoch a dimenziách.

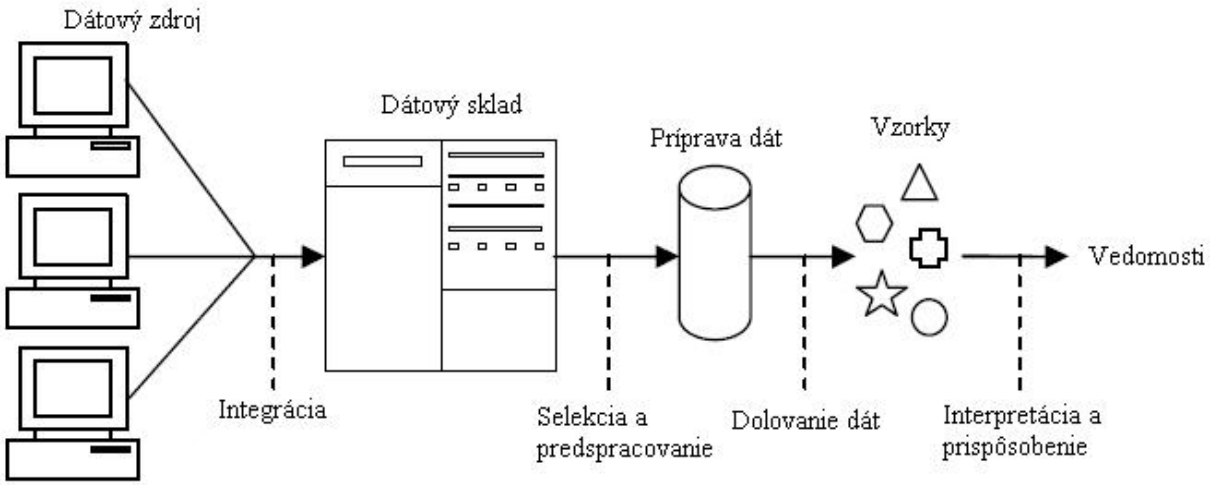
Fakty sú numerické merné jednotky. Tabuľka faktov je spravidla najväčšia tabuľka v databáze a obsahuje veľký objem dát. Niektoré jednoduchšie databázy zvyčajne obsahujú len jednu tabuľku faktov, iné môžu obsahovať viaceré tabuľky faktov. Prvotné fakty sa môžu kombinovať alebo vypočítať pomocou iných faktov a vytvoriť tak merné jednotky. Merné jednotky sa môžu uložiť v tabuľke faktov, prípadne vyvolať, ak je to nevyhnutné, na účely vykazovania.

Dimenzie obsahujú logicky alebo organizačne hierarchicky usporiadané údaje. Sú to vlastne textové popisy. Tabuľky dimenzií sú zvyčajne menšie ako tabuľky faktov a dáta v nich sa nemenia tak často. Tabuľky dimenzií vysvetľujú všetky „prečo“ a „ako“ pokiaľ ide napr. o obchodovanie a transakcie prvkov. Kým dimenzie vo všeobecnosti obsahujú relatívne stabilné dáta, napr. dimenzie zákazníkov sa aktualizujú častejšie. (1), (3), (7)

3. Získavanie znalostí z databáz

Získavanie znalostí v databázach (KDD) je proces identifikácie validity, originality, užitočnosti a pochopiteľnosti vzoriek údajov z veľkých databáz. Dolovanie dát je jadrom procesu dobývania znalostí.

Získavanie znalostí možno tiež definovať ako netriviálnu extrakciu predtým neznámych a potencionálne užitočných informácií z dát. Dolovanie dát je len jedným procesom dobývania znalostí.



Obrázok 2: Proces získavania znalostí

Dáta prichádzajú z viacerých zdrojov a sú integrované a vložené do spoločného dátového skladu. Časť z nich je potom predpracovaná do štandardného formátu. Na tieto predpripravené dáta je potom aplikovaný algoritmus dolovania dát, ktorý produkuje výstup vo

forme pravidiel alebo ďalšieho druhu vzoriek. Tie sú potom interpretované ako nové a potencionálne užitočné vedomosti (informácie). (4),(5)

3.1. Metódy získavania znalostí z databáz

K najpoužívanejším metódam získavania znalostí z databáz patrí metóda 5A, SEMMA a CRISP-DM.

Metóda 5A zahrňuje týchto päť krokov:

- Assess – posúdenie potrieb projektu
- Access – zhromaždenie potrebných dát
- Analyze - prevedenie analýz
- Act - premena znalostí na akčné znalosti
- Automate - prevedenie výsledkov analýzy do praxe

Metóda SEMMA sa tiež skladá z piatich krokov:

- Sample - vyberanie vhodných objektov
- Explore - vizuálna explorácia a redukcia dát
- Modify - zoskupovanie objektov a hodnôt atribútov, dátovej transformácie
- Model - analýza dát
- Assess - porovnanie modelov a interpretácia

Metóda CRISP-DM (CRoss-Industry Standard Process for Data Mining) sa snaží nájsť univerzálne použiteľný postup pre dolovanie dát. Metodika CRISP-DM uvádza nasledujúce etapy:

- porozumenie problematike (Business Understanding)
- porozumenie dátam (Data Understanding)
- príprava dát (Data Preparation)
- modelovanie (Modeling)
- vyhodnotenie výsledkov (Evaluation)
- využitie výsledkov (Deployment)

Terminológia uvádzaná v rámci tejto metodiky tvorí v obore KDD už štandard, aspoň vo svojej anglickej jazykové mutácii.

4. Budúce smerovanie výskumu

Po stanovení oblasti v riadení procesov je potrebné vybrať vhodnú metódu KDD. Veľmi dôležitou súčasťou KDD je dolovanie dát. Dolovanie dát je proces objavovania zmysluplných nových korelácií, vzoriek a trendov na veľkých množstvách dát uložených v dátových skladoch, používanie vzoriek rozpoznávacích technológií založených na štatistických a matematických technikách. (6)

Metódy dolovania dát využívajú niektoré štatistické metódy, hlavne koreláciu, lineárnu a logistickú regresiu, diskriminačnú analýzu a metódy predpovedania. Zložitejšie postupy sú niekedy realizované pomocou neurónových sietí, fuzzy logiky (8), (9) a genetických algoritmov.

Vhodným výberom techník KDD, vrátane techník dolovania dát je možné zlepšiť riadiaci proces predpovedaním jeho stavov, znížiť náklady alebo zlepšiť efektivitu a stabilitu riadiaceho procesu.

5. Zoznam bibliografických odkazov

(1) Inmon William: Building The Data Warehouse. Wiley Publishing, 2005, 543 s., ISBN 0-7645-9944-5

(2) Bandyopadhyay Sanghamitra, et. al: Advanced Methods for Knowledge Discovery from Complex Data. Springer, 2005, 369 s., ISBN 1-85233-989-6

(3) Ľuboslav Lacko: Business Inteligence v SQL Serveru 2005. Computer Press, 2006, 391 s., ISBN 8025111105

(4) Fayyad Usama, et. al: From Data Mining to Knowledge Discovery in Databases. AI-Magazine, 1996, dostupné [on-line] na http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf

(5) Data Mining and Knowledge Discovery Handbook. Editor Maimon Oded a Rokach Lior, Springer, 2005, 1383 s., ISBN 0-387-24435-2

(6) Olson David – Delen Dursun: Advanced Data Minig Techniques. Springer, 2008, 180 s., ISBN 978-3-540-76916-3

(7) Tanuška P., Verschelde, W., Kopček M.: The proposal of Data Warehouse test scenario. Proceedings of III. ECUMICT, 2008 Gent Belgium, ISBN 9-78908082-553-6

(8) Bohacik, J.: Induction by fuzzy attribute elimination. Journal of Information, Control and Management Systems, Vol. 5, No. 2, 2007, pp. 291-301, ISSN 1336-1716

(9) Bohacik, J., Juhola, M., Levashenko, V.: Fuzzy IF-THEN rule induction with cumulative information estimations applied to real-world data. Acta Electrotechnica et Informatica, Vol. 8, No. 4, 2008, pp. 24-29, ISSN 1335-8243

6. Adresa autora:

Andrej Trnka, Ing.
Katedra aplikovanej informatiky FPV UCM
Nám. J. Herdu 2
917 01 Trnava
andrej.trnka@ucm.sk

Tento príspevok bol podporovaný grantovou agentúrou VEGA v rámci projektu číslo 1/4078/07, za čo autor vyslovuje poďakovanie.