ASISTAREA INFORMATICĂ A DECIZIEI PRIN DEMERSUL IDENTIFICĂRII DE CUNOŞTINȚE ÎN BAZELE DE DATE - KDD (KNOWLEDGE DISCOVERY IN DATABASES)

PROF.UNIV.DR. IONESCU BOGDAN
Academia de Studii Economice Bucuresti, Adresa de corespondenta: Str 9 Mai nr. 5, bl. 36, sc.B, et. 2, ap. 25, sector 6, Bucureşti, tel/fax 0212213446, mobil 0722 593695, e-mail bionescu@gmail.com
CONF.UNIV.DR. MIHAI FLORIN
Academia de Studii Economice Bucuresti, Adresa de corespondenta: str. Vintila Mihailescu 16, Bloc 70, ap 11, sector 6, Bucuresti, tel 0724255333, email fmihai@gmail.com

L'assistance informatique de la décision par la démarche de découverte des connaissances dans les bases des données.
La démarche KDD (découverte des connaissances contenues dans les bases des données) représente un domaine de la technologie d'information qui combine les approches des systèmes informatiques décisionnels, de reconnaître des modèles, d'analyse des données, de présentation de l'information, pour extraire automatiquement les relations, des concepts et des connaissances contenues dans les grandes bases des données. Le processus KDD met en évidence et interprète les schémas des connaissances existantes dans les données.

KDD este un domeniu emergent care combină tehnicile "maşinilor de învățare", de recunoaştere a modelelor, de statistică, de vizualizare, pentru a extrage în mod automat inter-relațiile de concepte şi module, conținute în marile baze de date. Sarcina de bază a tehnologiilor KDD este deci aceea *de a extrage cunoştințe, plecând de la nivelurile cele mai de jos ale bazelor de date*. KDD interpretează schemele de cunoştințe existente în bazele de date.

Figura următoare furnizează o idee despre locul acestor ansambluri față de o bază de date.

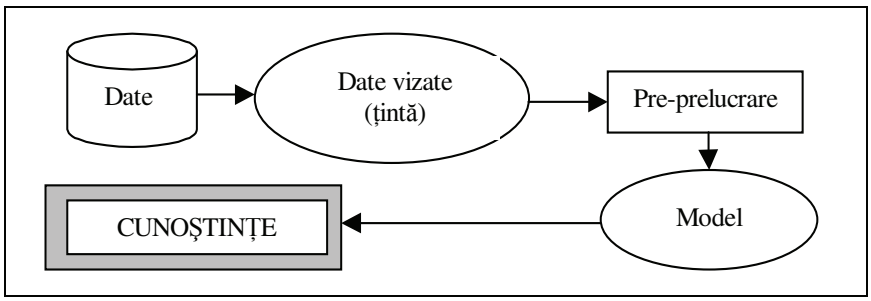

Figura 1. Legătura dintre date, modele, cunoştințe

Cunoştințele extrase sunt utilizate ca instrumente de asistare a deciziei, de exemplu, pentru operațiile de previziune şi clasificare, pentru a rezuma conținutul unei baze de date sau pentru explicarea fenomenelor şi tendințelor observate. Utilizarea sistemelor KDD permite *degrevarea decidenților de problema analizei manuale a marilor cantități de date din*

bazele de date. Aceste sisteme se dezvoltă şi actualmente încep a fi implementate în întreprinderi, în mod particular, în domeniul finanţelor, al diagnosticului, al analizei pieţelor, al detectării fraudelor, etc.

Facilităţile oferite de KDD devin extrem de importante în condiţiile creşterii continue a masei de date, a informaţiilor, a cunoştinţelor disponibile. Astfel, capacitatea şi rapiditatea procesului de introducere a datelor creşte tot mai mult datorită noilor tehnologii, în timp ce capacitatea de analiză a acestor date şi informaţii nu se dezvoltă în aceeaşi proporţie. Deci, se produce în mod inevitabil o sub-optimizare a ansamblului de resurse informaţionale. KDD îşi aduce o contribuţie majoră la soluţionarea acestei contradicţii pentru optimizarea sistemului informaţional.

Expresia KDD poate admite două interpretări: "KD in Data" şi "KD in Databases". Pentru a surmonta dificultatea de a gestiona un imens volum de informaţii, există două abordări posibile:

- reimplementarea algoritmilor de învăţare pentru a face operaţionale bazele de date, în scopul manipulării unui volum mare de date;
- reducerea volumului de date şi de atribute.

Acest ultim punct implică o prelucrare anterioară a datelor şi arată importanţa descoperirii de cunoştinţe în date (KD in Data). În acest caz, pot interveni elemente statistice, precum şi alte tehnici cum ar fi, avizul experţilor, existenţa şi calitatea cunoştinţelor în domeniu.

Pentru a avea maximum de beneficii din tehnologia KDD, o strategie mai completă este aceea de a utiliza cu preponderenţă descoperiri în date (KD in Data) şi nu în bazele de date (in Databases). Dezvoltarea abordărilor specifice KDD necesită încă timp pentru o cercetare aprofundată. Astfel de abordări, vor constitui o provocare pentru statisticieni, cercetători şi pentru informaticienii specialişti în domeniul bazelor de date. Procesul de pre-prelucrare este un instrument pentru abordări KDD aplicative, în particular, reducând cantitatea de date şi atributele aferente bazei de date.

Nu există o metodologie generală a tehnologiei KDD, aplicabilă tuturor cazurilor şi domeniilor, dar există o anumită experienţă ce poate permite definirea unui oarecare număr de etape generale, pentru a lăsa o marjă de manevră în funcţie de condiţiile în care se aplică procedura de descoperire de cunoştinţe.

Aplicaţiile de descoperire de cunoştinţe sunt aplicaţii ce posedă următoarele *caracteristici*:

- Cunoştinţele codate, ascunse în aplicaţii, sunt în sens larg extrase din datele întreprinderii. Aceste cunoştinţe sunt în general validate şi îmbogăţite prin avizul dat de experţii în domeniu;
- Rezultatul produs de către aplicaţii este reprelucrabil. Acest lucru este posibil numai dacă există un vocabular al întreprinderii în măsură să "înţeleagă" şi să poată fi utilizat direct, fără o altă transformare pentru sarcinile de gestiune;
- Cunoştinţele codate (ascunse) în aplicaţii, pot fi menţinute şi pot fi perfecţionate de utilizatorii întreprinderii, de o aşa manieră încât acestea pot fi modificate fără ca aplicaţia să fie recompilată.

Aplicaţiile ce au ca obiect descoperirea de cunoştinţe în date sau în bazele de date sunt concepute într-o primă abordare pentru utilizatorii informaţiei întreprinderii. Aceste

aplicații trebuie să încurajeze utilizarea decizională a informației şi, mai ales, trebuie să expliciteze reprezentările cunoştinţelor înconjurătoare.

Iată câteva probleme generale, ce trebuie luate în considerare în tehnologia KDD:

a) Construirea unui Data Warehouse

Se cunoaşte faptul că în majoritatea întreprinderilor, datele informatice se găsesc dispersate în diferite sisteme operaționale. Aceste sisteme au fost concepute pentru a prelucra şi actualiza date, dar această situație ce poartă amprenta eterogenității, nu este practic cea mai bună soluție pentru aplicațiile KDD. Pentru a dezvolta aplicațiile de "descoperire de cunoştințe", o întreprindere trebuie să-şi transpună datele existente din sistemele informatice operaționale într-un model de date unic, coerent şi logic care să acopere ansamblul necesităților informaționale. Acest sistem consolidat constituie un veritabil Data Warehouse. A construi un Data Warehouse reprezintă o etapă de pre-prelucrare necesară dezvoltării aplicațiilor ce au ca obiect "descoperirea de cunoştințe".

b) Descoperirea de cunoştințe

Această activitate începe cu înțelegerea sarcinilor de îndeplinit pentru a obține rezultatele aşteptate şi cu înțelegerea universului organizației în sensul statistic al termenului, plecând de la datele care "populează", un Data Warehouse. Bazându-se pe această înțelegere, instrumentele de clasificare şi vizualizare pot fi utile pentru a segmenta acest univers în subansambluri. Acest proces de segmentare este urmat de selectarea variabilelor independente pentru a construi şi dezvolta un model, după care se pune problema alegerii metodologiilor de modelare. Pentru acest lucru, trebuie să se țină cont de anumiți factori cum ar fi: tipul datelor prelucrate şi distribuția variabilelor independente. Aceşti factori trebuie să garanteze punerea în aplicare a acestor metodologii de modelare. La sfârşit, se dezvoltă un model, se testează şi se evaluează funcționalitatea sa.

c) Proiectarea unei "ieşiri reprelucrabile"

Această activitate este una de post-prelucrare, ce vine să răspundă la următoarea întrebare: "care vor fi rezultatele procesului de descoperire de cunoştințe?". Printre rezultatele reprelucrabile, ar trebui să se găsească "cadre" de informație (utile aplicațiilor de descoperire de cunoştințe pentru analize ulterioare), motoare ale bazelor de date active (utile pentru generarea rapoartelor asupra operațiunilor cheie ale întreprinderii) şi agenți (utili pentru implementarea normală a operațiunilor comerciale).

d) Actualizarea procesului de descoperire de cunoştințe

Pentru a realiza menținerea în exploatare a tehnologiei KDD, precum şi a criteriului de perfecționare a aplicațiilor ce au la bază descoperirea de cunoştințe, se remarcă aplicațiile de analiză denumite generic "post-hoc". În cazul acestor aplicații cunoştințele sunt reprezentate sub o formă cauzală abstractă, utilizatorul interacționând cu aplicația prin procedee de modificare a datelor. Astfel, prin procedee de actualizare, modelul este recalculat, pentru că acesta se bazează pe noi date.

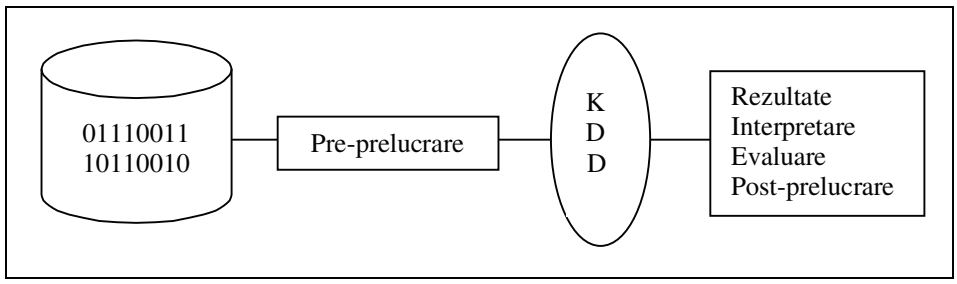

Figura 2. Procese de pre-prelucrare şi post-prelucrare

În general, se poate spune că trebuie ţinut cont de următoarele elemente atunci când se proiectează aplicaţii KDD:

- discuţii ample cu utilizatorul pentru a se descoperi natura şi finalitatea aplicaţiei. Etapa cea mai importantă în acest proces o constituie selectarea unei variabile dependente care furnizează într-adevăr pentru întreprindere informaţia prelucrabilă. În paralel, se regăseşte activitatea de achiziţie a tuturor variabilelor (din coloanele bazelor de date) pentru a consolida modelul. Câteva tipuri de probleme pot avea nevoie de numai 50 de variabile, în timp ce altele au nevoie de 500 sau 1000 de variabile;
- instrumentul KDD trebuie să fie capabil să prelucreze un mare volum de date, adăugând noi date la variabilele existente (problemă de consolidare). Câteva aplicaţii KDD pot genera rezultate utile, numai exploatând zeci de mii de înregistrări, în timp ce majoritatea aplicaţiilor exploatează curent sute de mii până la milioane de înregistrări.
- instrumentul KDD trebuie să fie uşor de utilizat. Aceasta înseamnă că interfaţa trebuie să furnizeze view-urile relaţiilor dintre date, calitatea previziunii şi a modelelor, într-un mod direct şi intuitiv. Ideal ar fi ca utilizatorii să nu aibă nevoie de cursuri de formare în acest domeniu, pentru a învăţa relaţiile între tehnologiile informaticii decizionale sau pentru a învăţa cum funcţionează interfeţele. Trebuie reflectat asupra faptului că cei mai mulţi beneficiari ai KDD sunt decidenţi şi că aceştia nu au timp de investit pentru a afla cum funcţionează noua tehnologie a informaticii decizionale;
- instrumentul KDD trebuie să manipuleze variabile numerice şi categoriale, chiar dacă datele sunt răspândite în diverse surse sau chiar lipsesc, iar relaţiile dintre date sunt adesea non-lineare.

Pentru ca tehnologia descoperirii de cunoştinţe să fie capabilă să extragă informaţia din marile baze de date, trebuie ca aceasta să interacţioneze îndeosebi cu sistemul de gestiune al bazelor de date, decât a obliga utilizatorul să-şi extragă datele din aceste SGBD-uri şi a le menţine în afara bazelor de date. Această interacţiune va facilita gestionarea datelor. Dacă mai mulţi utilizatori doresc să consulte în acelaşi moment bazele de date, aceştia riscă să nu aibă acces la informaţiile dorite. Din păcate, multe dintre sistemele de gestiune a bazelor de date nu sunt încă pregătite să rezolve această problemă.

Printre altele, în materie de KDD poate fi relevată următoarea contradicţie: pe de o parte, se asistă la creşterea în popularitate a descoperirii de cunoştinţe în bazele de date, iar pe de altă parte, există totuşi puţine aplicaţii care au fost încununate de succes. Există mai multe raţiuni care explică acest fapt:

- succesul depinde, în primul rând, de *alegerea unei aplicaţii adecvate*. O astfel de aplicaţie trebuie să necesite descoperirea de rezultate care se ştiu extrase din date şi ale căror prelucrări sunt cunoscute;
- succesul depinde în egală măsură de *conlucrarea aplicaţiilor informatice existente* şi *operaţionale în întreprindere* şi *metodele proprii descoperirii de cunoştinţe*. În acest sens, există mai multe metode şi abordări, dar pentru moment nici una nu se detaşează net;
- *tranziţia de la sistemul de explorare KDD la aplicaţia obişnuită*, face faţă aceloraşi probleme ca de altfel toate transferurile de expertiză, de la cercetare la aplicaţiile curente;
- publicaţiile referitoare la KDD există, dar ele se adresează unui mic număr de iniţiaţi.

Utilizarea aplicaţiilor KDD este eficace în situaţiile în care volumul datelor necesare pentru a extrage o soluţie este mic sau poate fi redus utilizând diferite tehnici statistice. În general, performanţa algoritmilor ce reies din aceste aplicaţii asupra ansamblului de date introduse este non-lineară. Pentru a se putea servi de aceste aplicaţii, utilizatorii trebuie să stăpânească caracteristicile algoritmilor utilizaţi sau să aibă cunoştinţe temeinice asupra tehnicilor statistice. Aceasta reprezintă o limită importantă de care trebuie să se ţină seama şi care trebuie rezolvată printr-o tehnologie ce permite aplicaţii mai uşor de utilizat şi care beneficiază de interfeţe mult mai conviviale.

Noile tehnologii de asistare a deciziei nu vin să înlocuiască exhaustiv sistemele clasice de asistare a deciziei, ci vin să le completeze, pentru ca decizia (în speţă cea financiar–contabilă) – să poată considera şi alte elemente "ascunse în date". Aceste tehnologii ale informaticii decizionale, pot conlucra atât cu sistemele informatice clasice existente în întreprindere – pe care le exploatează , dar mai ales cu tehnologiile inteligenţei artificiale care la rândul lor utilizează expertizele furnizate de noile abordări ale informaticii decizionale.

BIBLIOGRAFIE:

1. B. IONESCU, Sisteme de asistare a deciziei financiar contabile, Ed. InfoMega, Bucuresti, 2002
2. F. MIHAI, Sisteme informationale financiar – contabile, Ed. InfoMega, Bucuresti, 2004
3. KLEIN M., METHLIE L.B., Knowledge-based Decision Support Systems with applications in business, John Wiley&Sons, New York, 1999
4. PALLER Allan, LASKA R., The EIS book: Information System for top managers, Journal of Information Systems Management, Vol. 7, No.4, 1998
5. POWER D.J., Justifying a Data Warehouse project, http://dss.cba.uni.edu/papers/dsscontext/index.html