MANAGEMENT INTERCULTURAL

Volumul VI, Numărul 10 / 2004

ISSN 1454-9980

Procesul de data mining: descoperirea cunoştinţelor ascunse

(pag. 19-24)

Bogdan ANASTASIEI

Volume VI, Issue 1 (10) / 2004

Cross-cultural Management Journal

PROCESUL DE *DATA MINING:* DESCOPERIREA CUNOŞTINŢELOR ASCUNSE

Bogdan ANASTASIEI

"Avem foarte multe date colectate în firmă, ce să facem cu ele acum?" Această problemă a devenit una obişnuită în multe organizaţii mari. Informaţia digitală este ieftin de obţinut si relativ ieftin de stocat. Dar care este scopul stocării unei cantităţi de date atât de mari? În afara argumentului legat de modalităţile convenabile de stocare a datelor în format electronic, mai există şi altul: firmele colectează date deoarece managerii "simt" că aceste date reprezintă un activ valoros, care ar putea fi folosit la un moment dat. În institutele de cercetare, datele se referă la observaţii asupra fenomenelor aflate sub studiu, culese cu atenţie. În organizaţiile economice, datele şi informaţiile se referă la pieţe, concurenţi, furnizori, distribuitori, clienţi, precum si la informaţii interne legate de procesele de producţie, organizarea muncii, depozitare etc.

Pentru multe firme din Occident, procesul de explorare şi analiză a datelor nu este unul nou. Dar programele de *data mining* permit realizarea acestor analize mult mai rapid şi, potenţial, cu o eficienţă sporită. În multe cazuri, tehnicile de explorare a datelor permit ca datele colectate pentru un anumit scop să poată fi folosite în multe alte scopuri.

Ca urmare a perfecţionării modalităţilor de colectare şi stocare a datelor, foarte multe companii mari s-au trezit deodată că stau pe un munte de date cărora ar trebui să le găsească o utilizare. De exemplu, companiile furnizoare de cărţi de credite înregistrează permanent datele referitoare la tranzacţiile comerciale, supermarketurile înregistrau date privind cumpărăturile, utilizarea cupoanelor de reduceri ş.a.m.d. Apariţia necesităţii de a valorifica toate aceste date era doar o chestiune de timp.

Prin *data mining* se înţelege procesul de extragere automată a unor informaţii cu caracter predictiv din marile baze de date. Ea poate prezice tendinţele viitoare şi identifica comportamente care le-au scăpat responsabililor din organizaţie.

Datele brute sunt rareori folositoare în forma în care se găsesc. Valoarea lor reală este dată de posibilitatea de a extrage din ele informaţii care să ajute la fundamentarea deciziilor sau la înţelegerea fenomenelor care guvernează viaţa organizaţiei. În mod tradiţional, analiza datelor era un proces strict manual. Unul sau mai mulţi analişti deveneau foarte familiari cu datele şi furnizau periodic rapoarte, cu ajutorul metodelor statistice. O asemenea abordare devine totuşi dificil de pus în practică, pe măsură ce volumul de date creşte până la dimensiuni uriaşe. Cine ar putea "înţelege" o bază de date cu câteva milioane de obiecte, fiecare având câteva zeci de câmpuri? Pentru a complica situaţia, volumul de date creşte într-un ritm atât de accelerat încât analiza manuală (chiar şi atunci când este posibilă) nu mai poate ţine pasul.

Comunitatea de cercetători şi practicieni interesaţi de problema automatizării analizei datelor a crescut rapid. Două sunt denumirile sub care este cuprins acest proces de automatizare: "knowledge discovery in databases" (KDD) şi "data mining", pe care noi o vom mai numi aici şi "explorarea datelor". Primul workshop pe tema KDD a fost ţinut în 1989; el s-a transformat ulterior într-o conferinţă internaţională care reuneşte anual peste 500 de participanţi.

Explorarea datelor ar putea fi de asemenea definită ca "procesul de căutarea a unui *pattern*

(model) în interiorul datelor". Ea utilizează metode sofisticate de analiză statistică. Odată găsită, informaţia ascunsă în spatele datelor brute trebuie prezentată într-o formă accesibilă cu ajutorul rapoartelor, tabelelor, graficelor etc.

Statistica este pilonul central al întregii activităţi de explorare a bazelor de date. Atât în ceea ce priveşte validarea ipotezelor, cât şi în ceea ce priveşte analiza exploratorie a datelor şi realizarea predicţiilor, statistica are o importanţă fundamentală.

Calitatea datelor este critică pentru obţinerea unor rezultate consistente din analiza datelor. Primul pas în KDD şi în *data mining* este construirea unui sistem de culegere a datelor într-o formă cât mai precisă şi fidelă. Datele primare de calitate slabă pot compromite întregul proces.

Sistemul cu ajutorul căruia se realizează în fiecare zi activităţile din organizaţie precum preluarea comenzilor, înregistrările contabile, gestiunea stocurilor etc. poartă numele de "sistem operaţional", iar informaţiile pe care acest sistem le creează poartă denumirea de "date operaţionale". Explorarea datelor este folosită pentru a da consistenţă şi substanţa acestor date operaţionale, care sunt date brute, primare. Este un proces de "descoperire", în urma căruia ies la iveală informaţii pe care datele brute nu le pot oferi în forma actuală. Nu sunt utilizate rapoarte cu o structură prestabilită, ci este permisă decidentului crearea de interogări a bazei de date pornind de la informaţiile de care are nevoie în momentul respectiv.

Să luăm un exemplu de aplicare a procesului de *data mining* în vânzări. Utilizatorul (managerul de vânzări) poate începe prin a cere bazei de date informaţii cu privire la vânzările totale pe trimestrul curent şi pe trimestrul anterior, precum şi un calcul al diferenţei şi a procentajului de creştere sau descreştere a vânzărilor. El află, de exemplu, că vânzările au crescut, per total, cu 3,5% de la trimestru la trimestru. Mai departe, el doreşte să afle cum stă evoluţia vânzărilor pe fiecare din cele 42 de judeţe, şi cere bazei de date un raport asemănător, numai că acum defalcat pe judeţe. Primind acest raport, el află că vânzările au crescut în majoritatea judeţelor, dar au scăzut în Bacău şi Vrancea.

Pentru a afla mai mult, el cere situaţia vânzărilor pentru fiecare din cei 10 distribuitori din judeţul Bacău, unul din judeţele cu probleme. Astfel, află că în acest judeţ, 8 din 10 distribuitori au înregistrat scăderi destul de drastice ale vânzărilor. Mai departe, el cere pentru fiecare din aceşti distribuitori, o situaţie a evoluţiei vânzărilor pe produse şi observă că declinul cel mai accentuat se înregistrează la produsul X.

Urmărind situaţia vânzărilor pentru zona Vrancea, utilizatorul descoperă acelaşi lucru: vânzările la produsul X scad la majoritatea distribuitorilor din acest judeţ. Studiind vânzările produsului X şi în judeţele unde se înregistrează creşteri de ansamblu ale cifrei de afaceri, managerul află că în cazul acestui produs, creşterile sunt foarte reduse (aproape de zero), iar în multe cazuri (cca. 30% din totalul judeţelor) s-au produs chiar scăderi uşoare.

Concluzia analizei: produsul X pare a fi un produs-problemă. Momentan, problema se manifestă mai acut în două zone, dar s-ar putea extinde în viitor. Aşadar, deşi per ansamblu vânzările firmei au crescut, există totuşi o problemă importantă, care a fost identificată cu ajutorul procesului de *data mining*.

Iată câteva alte aplicaţii posibile ale procesului de *data mining* în organizaţiile economice (şi nu numai).

- în marketing, la realizarea profilului consumatorilor. Caracteristicile clienţilor buni (celor care cumpără de la firmă, regulat şi în cantităţi mari) sunt utilizate ca variabile de predicţie. Ele îi vor ajuta pe marketeri să-i "ţintească" pe noii clienţi. Cu ajutorul explorării putem identifica, în baza de date cu clienţi, o serie de modele care pot fi aplicate ca criterii de selectare pentru o bază de date cu clienţi potenţiali. Acei clienţi potenţiali, care se aseamănă cel mai mult cu "clientul ideal", vor constitui cei mai buni candidaţi pentru o campanie de marketing direct, de exemplu. Este de aşteptat ca beneficiile oferite de produsul firmei, să fie interesante şi pentru ei. Acest fapt permite eficientizarea campaniilor de marketing, prin direcţionarea lor către acei potenţiali clienţi care au cea mai mare probabilitate de cumpărare;
- explorarea datelor îi poate ajuta pe detailişti să înţeleagă cum arată "coşul de produse" al unui cumpărător individual (ce tip de produse şi ce mărci sunt achiziţionate simultan). Ei pot afla astfel ce produse trebuie expuse pe rafturi şi cum trebuie

expuse. Explorarea poate de asemenea ajuta la măsurarea eficienţei campaniilor promoţionale ale magazinului;
- o altă utilizare obişnuită a explorării datelor o întâlnim în gestiunea relaţiilor cu clienţii. Determinând caracteristicile clienţilor pentru care este foarte mare probabilitatea de a ne părăsi în favoarea unui client, compania poate pune în practică acţiuni de reţinere a clienţilor, ştiut fiind că reţinerea clienţilor vechi costă mult mai puţin decât atragerea de noi clienţi;
- detectarea fraudelor este un alt câmp de aplicaţie pentru *data mining*, de care sunt interesate în special companiile de telecomunicaţii, organizaţiile care emit cărţi de credit, bursele, agenţiile guvernamentale. Suma totală pierdută din cauza fraudelor este enormă. Cu ajutorul explorării datelor, organizaţiile pot identifica tranzacţiile potenţial frauduloase şi stopa pierderile până nu este prea târziu;
- organizaţiile financiare utilizează explorarea datelor în scopul determinării caracteristicilor pieţei şi a diferitelor sectoare de activitate, prezicând astfel comportamentul viitor al titlurilor financiare ale diferitelor companii;
- o altă aplicaţie este cea din domeniul medical: explorarea datelor poate prezice eficienţa intervenţiilor chirurgicale, analizelor medicale, administrării unor medicaţii, serviciilor medicale ş.a.m.d.

Bazele de date utilizate pentru acest proces de explorare trebuie să fie deosebit de flexibile, deoarece utilizatorul poate face explorarea într-o manieră ad-hoc, fără a urma standarde prestabilite. În exemplul nostru anterior, managerul şi-a structurat interogarea pe direcţia judeţ-distributori-produs, dar altcineva ar putea cere foarte bine un raport pe direcţia produs-distribuitori-judeţ sau produs-judeţ-distribuitori. Baza de date trebuie să cuprindă foarte multe referinţe încrucişate, pentru a putea asocia cu uşurinţă câmpurile şi extrage rapid rapoarte rezumative. O altă cerinţă importantă este aceea ca datele să fie cât mia "condensate" cu putinţă, astfel încât timpul necesar obţinerii răspunsului la o interogare să fie cât mai redus posibil (câteva sutimi de secundă). Dacă timpul necesar obţinerii unui raport este prea mare (de ordinul secundelor), utilitatea procesului este pusă sub semnul întrebării.

Caracteristicile cheie ale unui bun proces de *data mining* sunt:
- uşurinţa utilizării. Procesul trebuie să fie uşor de realizat şi să nu solicite o instruire prea îndelungată a utilizatorilor. Este important ca utilizatorul să se poată concentra asupra propriilor sale necesităţi şi probleme, asupra modului de formulare a interogărilor, şi nu asupra modului de utilizare a programului. Cu cât este mai dificilă rularea programului de *data mining*, cu atât este mai sigur că nu va fi folosit chiar de către persoanele care ar avea cea mai mare nevoie de el;
- accesibilitatea. Valoarea reală a explorării datelor se obţine oferind acest instrument celor care au nevoie de el, care ar avea de aflat cele mai multe lucruri în urma analizei datelor. Managerii de vânzări, de exemplu, îl vor folosi pentru a evalua eficienţa agenţilor de vânzări şi a distribuitorilor, pe fiecare tip de client în parte. Managerii de achiziţii îl vor folosi pentru a determina evoluţia stocurilor pe fiecare tip de material. Pentru contabili, ar putea fi un instrument util în construirea bugetelor şi a situaţiilor financiare, şi aşa mai departe;
- răspuns rapid. Este foarte important ca analiza şi furnizarea informaţiei să fie realizată în cel mai scurt timp posibil (de ordinul zecimilor sau sutimilor de secundă). Numai în acest fel capacităţile procesului de *data mining* vor fi folosite la întregul lor potenţial. Decidenţii vor fi mai puţin înclinaţi spre a utiliza programul atunci când ştiu că trebuie să aştepte secunde bune pentru a obţine rezultatul unei interogări, deoarece acest lucru ar însemna pentru ei o pierdere de timp importantă;
- informaţie la zi (actualizată). Folosirea unor date vechi de săptămâni sau luni reduce considerabil eficienţa şi utilitatea procesului. Cu cât datele sunt mai actuale, cu atât mai repede poate fi produs un plan de măsuri pentru rezolvarea problemelor sau exploatarea oportunităţilor identificate prin procesul de explorare a datelor. În exemplul nostru, combaterea declinului înregistrat la vânzările produsului X poate fi realizată cel mai eficient, dacă problema a fost descoperită chiar în momentul apariţiei ei şi cât mai devreme după acest moment.

Beneficiile procesului de explorare a datelor sunt următoarele:

- costul mai redus al procesării datelor. Atunci când decidenţii obţin datele necesare în urma unor analize făcute de ei înşişi, dispare necesitatea de a crea programe informatice specializate pentru obţinerea şi raportarea acestor date. Astfel, programatorii îşi pot concentra atenţia asupra altor probleme legate de operarea sistemelor din firmă;
- reducerea cantităţii de hârtie ocazionate de tipărirea rapoartelor şi situaţiilor. Marea majoritate a programelor informatice oferă managerilor sumele totale ce reflectă volumul activităţii din domeniul lor (de exemplu, vânzările totale). În cazul procesului de explorare a datelor, managerul poate începe prin a solicita aceste sume totale pe ecranul computerului, apoi poate cere defalcarea lor în funcţie de informaţia pe care doreşte să o primească. Odată ce deţine această informaţie, el va cere eventual tipărirea unui raport, foarte probabil mult mai util decât cel ce prezintă doar totalurile (justificând aşadar mai bine costul hârtiei pe care a fost tipărit);
- reducerea încărcării sistemului operaţional. Procesul de *data mining* utilizează de regulă baze de date care au fost realizate şi condensate fie offline, fie în afara orelor de program din firmă. Întrucât procesele de sortare, selectare şi raportare sunt mai uşoare pe aceste baze de date, încărcarea sistemului informatic al organizaţiei va fi substanţial redusă.

Exemplele din acest articol sunt exemple simple, care totuşi ne arată cât de util este a oferi decidenţilor posibilitatea de a privi datele din diverse unghiuri de vedere. Pe măsură ce întreprinderile îşi dau seama că procesul de *data minig* nu este unul extrem de complex şi de costisitor, tot mai multe îşi vor include achiziţionarea sistemelor de acest fel în planurile lor de dezvoltare viitoare.

Există trei clase de activităţi de *data mining*:

- descoperirea, care este procesul de căutare în bazele de date a unor modele ascunse, fără ipoteze şi idei preconcepute asupra naturii acestora. Cu alte cuvinte, iniţiativa de a căuta aceste modele îi aparţine programului, înainte ca utilizatorul să formuleze interogările;
- modelarea predictivă, care presupune utilizarea modelelor descoperite anterior în scopul prezicerii viitorului. Programul permite utilizatorului să introducă înregistrări care au unele câmpuri cu valoare necunoscută; valoarea acestora urmează a fi estimată de program tot cu ajutorul pattern-urilor identificate anterior;
- analiza extremelor, este procesul ce permite detectarea, pe baza modelelor identificate în prima etapă, a situaţiilor anormale sau neobişnuite ce caracterizează obiectele din baza de date. Pentru a descoperi cazurile neobişnuite, va trebui să se răspundă mai întâi la întrebările: “Ce înseamnă normalul? Care este regula?”. Cazurile care înregistrează o deviere, cu o anumita mărime, de la această regulă, vor fi considerate cazuri neobişnuite. Analiza extremelor merge aşadar mai departe decât descoperirea şi modelarea predictivă, ajutându-ne nu doar să obţinem aşa numitele “cunoştinţe obişnuite”, ci să identificam şi situaţiile anormale.

Putem vorbi de trei tipuri de activităţi de explorare a datelor care se desfăşoară într-o organizaţie: explorare episodică, explorare strategică şi explorare continuă.

Explorarea episodică presupune efectuarea analizei datelor cu ocazia unui anumit eveniment: o campanie de marketing, de exemplu. Se pot analiza datele legate de această campanie pentru a se înţelege o serie de relaţii de cauzalitate (de exemplu, relaţia dintre cheltuielile pentru publicitate şi volumul vânzărilor) şi se poate folosi această analiză pentru a planifica şi previziona rezultatele viitoarelor campanii de marketing. De regulă, analiştii sunt cei care realizează activităţi de *data mining* episodice.

Explorarea strategică presupune luarea în considerare a unui set de date mai extins, pentru a dobândi o înţelegere de ansamblu asupra unor parametri care caracterizează activitatea firmei (de exemplu, profitabilitatea). În acest caz, explorarea strategică poate răspunde la întrebări de genul: “Care sunt principalele noastre surse de profit?” sau “Care este relaţia dintre portofoliul nostru de produse şi segmentele noastre de piaţă?”.

În cazul explorării continue, obiectivul este acela de a înţelege ce schimbări au avut loc în organizaţie şi în mediul ei extern pe o anumită perioadă de timp şi care sunt cei mai importanţi factori care au influenţat acele schimbări. Managerul s-ar putea întreba, de exemplu: "cum au evoluat vânzările în ultimul semestru, pe total şi pe tipuri de produse?" sau "cum s-a modificat nivelul de satisfacţie a consumatorilor noştri în ultimele luni şi din ce motive?".

Procesul de explorare a datelor are cinci componente principale, care sunt de fapt grupe de tehnici de analiză a datelor. Iată care sunt acestea:

- clasificarea (sau *clustering*-ul);
- asocierea;
- analiza secvenţială;
- reţelele neuronale;
- arborii de decizie

În ceea ce urmează le vom discuta pe fiecare din ele pe scurt.

Clasificarea analizează un set de date şi stabileşte un ansamblu de reguli pe baza cărora vor fi grupate datele obţinute în viitor. Software-ul de *data mining* identifică automat clasele (sau "ciorchinii"), studiind pattern-ul datelor existente. Odată ce au fost generate clasele se poate stabili, pe baza unor caracteristici precizate, cărei clase aparţine un anume obiect din baza de date studiată. De exemplu, o clasă poate fi un segment de piaţă. Atunci când firma are un client nou ea poate stabili, cunoscând caracteristicile acestuia, în ce segment se încadrează.

O regulă de asociere este o regulă care implică existenţa unor relaţii între diferite seturi de obiecte din baza de date. Procesul de asociere duce la descoperirea unor asemenea reguli, la diferite niveluri de abstracţiune, între obiectele bazei de date. De exemplu, se poate afla că scăderea vânzărilor este totdeauna însoţită de o serie de simptoame (evenimente din interiorul sau din afara organizaţiei) şi pe această bază se pot studia mai departe motivele care stau în spatele acestei asocieri.

Analiza secvenţială duce la descoperirea acelor evenimente care se petrec întotdeauna într-o anumită secvenţă. Ea studiază datele care apar în tranzacţii diferite (spre deosebire de asociaţie, care studiază relaţiile dintre datele ce apar în cadrul aceleiaşi tranzacţii). În urma analizei secvenţiale s-ar putea descoperi că majoritatea clienţilor, care în decursul unei săptămâni cumpără produsele A şi B, săptămâna viitoare cumpără produsele C şi D, de exemplu.

Pe măsură ce procesul de *data mining* devine mai obişnuit în organizaţii, reţelele neuronale şi arborii de decizie se bucură de tot mai multă consideraţie. Deşi reţelele neuronale sunt mai complexe în felul lor, utilizarea lor nu cere cunoştinţe de statistică prea multe şi prea avansate.

Reţelele neuronale utilizează un număr mare de parametri pentru a construi un model care preia şi combină un set de inputuri în scopul prognozării comportamentului unei variabile cantitative sau categoriale. Valoarea fiecărui nod al reţelei se calculează ca medie ponderată a valorilor nodurilor anterioare. Procesul de construire a modelului implică aflarea acelor ponderi care produc cele mai precise prognoze, utilizând date reale pentru a "antrena" reţeaua. Cea mai comună metodă de "antrenare" este compararea datelor calculate în reţea cu valorile corecte cunoscute. După fiecare comparare, ponderile sunt ajustate şi valorile sunt calculate din nou. După o perioadă îndelungată de asemenea comparări şi ajustări, o reţea neuronală devine de regulă un foarte bun predictor.

Arborii de decizie reprezintă o serie de reguli a căror aplicare ne conduce către o anumită clasă sau valoare din baza de date. De exemplu, să presupunem că o bancă vrea să-i clasifice pe solicitanţii de credite în două categorii: cei cu risc scăzut şi cei cu risc ridicat. Un arbore de decizie construit în acest scop ar putea arăta în felul următor:

- are solicitantul un venit mai mare de 50 000 de dolari pe an? Dacă da, următoare întrebare care se pune este: "Cât de mare este nivelul datoriei sale în prezent?". Dacă persoana respectivă este foarte îndatorată, ea prezintă un risc mare la creditare; dacă este puţin îndatorată, atunci prezintă un risc redus;
- dacă solicitantul are un venit mai mic de 50 000 de dolari pe an, întrebarea următoare este: "Care este vechimea sa în muncă?". Dacă este mai mare de 5 ani, se consideră că solicitantul prezintă un risc mic la creditare; dacă vechimea este mai mică de 5 ani, atunci este considerat un solicitant cu risc mare.

Arborii de decizie sunt folosiţi deseori, deoarece au un grad de precizie rezonabil şi, spre deosebire de reţelele neuronale, sunt mai uşor de înţeles şi necesită un timp de construire mai redus. Ar mai trebui precizat că atât arborii de decizie, cât şi reţelele neuronale pot fi utilizate pentru construirea unor modele de regresie.

Etapele unui proces de explorare a datelor (*data mining*) sunt în general următoarele:

- definirea problemei. Pentru o utilizare cât mai eficientă a tehnicilor de explorare trebuie mai întâi specificate foarte clar obiectivele studiului. De exemplu, în cazul unei campanii de *direct mailing*, obiective precum "creşterea numărului de răspunsuri", "creşterea ratei răspunsurilor" sau "diminuarea costului per răspuns" sunt obiective diferite, care vor necesita metode de analiză diferite. Definirea clară a obiectivelor ne va ajuta ulterior şi la măsurarea eficienţei procesului;
- selectarea tehnicii care va fi utilizată, pornind de la problema specifică de rezolvat;
- selecţia şi pregătirea datelor. Aceasta este etapa care consumă cel mai mult timp (între 50 şi 85 la sută din timpul total alocat proiectului). După cum spuneam anterior, calitatea rezultatului final depinde foarte mult de acurateţea datelor de intrare;
- construirea modelului, ţinând seama de tehnicile de analiză alese şi de natura datelor pe care le avem la dispoziţie;
- prezentarea rezultatelor finale, sub forma unor rapoarte conţinând text, tabele şi grafice;
- monitorizarea modelului, în vederea măsurării eficienţei şi utilităţii sale.