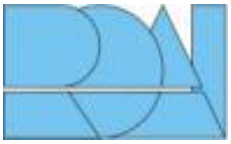
Mihai Istrate

UNIVERSITATEA DIN CRAIOVA



Centrul de Cercetare în Inteligență Aplicată – „Nicolae Țăndăreanu”
Research Center for Applied Intelligence – „Nicolae Țăndăreanu”
www.rcai.eu

Colecția « Computer Science »

Inițiată în 2001, sub egida Centrului de Cercetare în Inteligență Aplicată – „Nicolae Țăndăreanu”, colecția „Computer Science” reunește contribuții valoroase – publicații științifice de înaltă ținută, studii, teze de doctorat, etc. – continuând astfel tradiția publicării în volum separat a seriei 100 (lucrările conferinței anuale AIDC – Artificial Intelligence and Digital Communications), a seriei 200 (Computer Science Fundamentals) și a seriei 300 (Research Reports in Artificial Intelligence).

Rămânând fidelă misiunii sale de dezvoltare și promovare a cunoașterii în domeniul științei calculatoarelor și a tehnologiilor digitale, colecția „Computer Science” reprezintă o sursă reală de informare și își propune lărgirea spectrului publicistic prin dezvoltarea de noi serii tematice.

Propunerile pentru publicare se vor adresa comitetului științific la adresa: office@rcai.eu

Mihai Istrate

METODE DATA MINING APLICATE ÎN E-COMMERCE ȘI INTERNET

- MONOGRAFIE -

Editura UNIVERSITARIA
Craiova, 2021

Referenți științifici:
Prof.univ.dr. Daniela Dănciulescu
Lect.univ.dr. Claudiu-Ionuț Popirlan

Descrierea CIP a Bibliotecii Naţionale a României
ISTRATE, MIHAI
Metode data mining aplicate în e-commerce și internet / Mihai Istrate. - Craiova : Universitaria, 2021
Conţine bibliografie
ISBN 978-606-14-1705-6
004

LISTA DE FIGURI

LISTA DE TABELE

Capitolul 1

ANALIZA - DATA MINING

1.1 Elemente generale în DATA MINING

Încă de la începutul anilor 1990 se tot vorbeşte despre conceptul Data Mining (prescurtat DM), sau mai este întâlnit şi în literatura de specialitate română *minerit* în date, în foarte multe medii, pornind de la cele academice şi până la cele de afaceri sau medicale, îndeosebi.

Reprezentând o arie de cercetare cu o istorie destul de scurtă, nedepăşind faza *adolescenţei*, este încă disputată de câteva domenii ştiinţifice care o revendică.

Preluând afirmaţiile lui Pregibon -Research Scientist Google Inc. - şi anume că: *Conceptul Data Mining reprezintă un amestec de Statistică, IA (Inteligenţă Artificială) şi cercetare în baze de date*, sau faptul că anumiţi cercetători consideră Data Mining ca fiind *un cuvânt murdar în Statisticii* [122] - cel mai exact aceştia au fost statisticieni care nu considerau cu ceva timp în urmă Data Mining ca fiind ceva important pentru ei.

În ultimii ani, descoperirea de cunoştinţe şi instrumente de data mining au fost utilizate în principal în medii experimentale şi medii de cercetare. Acum suntem într-o etapă în cazul în care instrumentele sofisticate sunt în curs de dezvoltare rapid. Grupul Meta a estimat că dimensiunea pieţei pentru piata de exploatare a datelor a crescut de la cca. 50 milioane dolari în 1996 la cca. 800 milioane dolari prin 2000 [157].

În cele ce urmează vom detalia elemente fundamentale legate de conceptul Data Mining [98], cum ar fi:

- Ce este sau ce nu este conceptul Data Mining?
- De ce acest concept Data Mining?
- Cum se *sapă* sau *minereşte* în date?

- Probleme rezolvabile cu metode Data Mining.
- Despre modelare şi modele în conceptul Data Mining.
- Aplicaţii ale conceptului Data Mining.
- Exemplificarea terminologică a conceptului Data Mining.
- Confidenţialitatea datelor specifice.
- Numărul tot mai mare de baze de date de pe Web ce trebuie să fie exploatate pentru a extrage informaţii utile.

Data mining este procesul care prezintă diferite interogări şi extragerea unor informaţii utile, modele şi tendinţe (de cele mai multe ori necunoscute anterior). În esenţă, pentru multe organizaţii, obiectivele data mining includ îmbunătăţirea capacităţilor de marketing, detectarea modelelor anormale şi estimarea viitoare bazată pe experienţele trecute şi tendinţele actuale. Există în mod clar o nevoie destinată acestei tehnologii ([135], [156]).

Există cantităţi mari de date curente şi istorice. Prin urmare, bazele de date devin mai mari, tot mai dificil de a sprijini procesul decizional. În plus, datele ar putea fi din mai multe surse şi mai multe domenii. Există o nevoie clară de analiză a datele pentru a sprijini planificarea şi alte funcţii specifice unei întreprinderi.

Data mining a fost folosit până acum în afaceri de către organizaţii comerciale de succes în scopul de a obţine avantaje critice în competiţia lor. Se bănuieşte ca în viitorul apropiat acest instrument va fi folosit pentru prelucrarea bazelor de date uriaşe, ca de exemplu dosarele computerizate ale pacienţilor, la nivel naţional. De fapt, chiar şi în prezent, prin identificarea procedurilor medicale ce au tendinţa de a se grupa, prin data mining putem prezice care pacienţi vor folosi noile strategii de îngrijire a sănătăţii, putem defini modele de comportare ale pacienţilor de risc, putem identifica fraudele.

Conceptul de Data Mining [29] reprezintă *descoperirea şi analiza cunoştinţelor stocate în baze de date*, cu trei rădăcini generice, potrivit cărora a împrumutat tehnicile de lucru şi terminologia:

1. **Statistica** - reprezintă rădăcina cea mai lungă din Data Mining fără de care acest concept nu ar exista. Statistica clasică aduce în Data Mining unele tehnici definitorii rezumate în Analiza exploratorie a datelor (din denumirea ştinţifică a literaturii străine EDA = Exploratory data analysis), utilizată pentru a identifica relaţiile sistematice ale diferitelor variabile în momentul în care nu sunt infomaţii suficiente

asupra naturii acestora. Dintre tehnicile analizei exploratorie a datelor clasice care sunt utilizate în Data Mining, sunt detaliate:

(a) **metodele computaţionale:**

- *statistica descriptivă* (care face referire la repartiţii şi parametri);
- *statistici clasice* (se are în vedere media, mediana, deviaţia standard ş.a.);
- *corelaţii* (reprezentând tabele multiple de frecvenţă);
- *metode statistice multivariate* (mai exact metoda clustering, analiza factorilor, a funcţiei discriminant şi a corespondenţelor, arbori de clasificare şi de regresie, modele de regresie liniară şi neliniară, serii temporale ş.a);

(b) **vizualizarea datelor**, reprezintă metoda forte de explorare a datelor prin reprezentarea informaţiei în format vizual. Tehnicile de vizualizare sunt:

- histogramele;
- graficele rectangulare (anume box plots);
- graficele de împrăştiere (anume scatter plots);
- graficele de contur;
- graficele matriciale ş.a [162].

2. **Inteligenţa artificială (AI = Artificial Intelligence)** vine în ajutorul dezvoltării conceptului de Data Mining pe baza unor tehnici de procesare a informaţiei bazate pe modelul rationamentului uman. Învăţarea automată (conform literaturii de specialitate străină ML = machine learning) reprezintă o arie foarte de importantă a Inteligenţei Artificiale în raport cu dezvoltarea Data Mining, prin utilizarea tehnicilor care permit computerului să poată *învăţa* prin *antrenament*.

3. **Sisteme de baze de date** (conform literaturii de specilitate străină DBS = database systems), presupune cea de-a treia rădăcină a conceptului Data Mining, procurând materialul care trebuie *minerit* utilizând metodele amintite anterior.

Din perspectiva anumitor domenii necesitatea procesului de *minerit* al datelor poate fi rezumată astfel:

1. *Din punct de vedere economic* (afaceri-finanţe) - există un mare volum de date deja colectate în diverse domenii ca: date Web, e-commerce, super/hypermarket-uri, tranzacţii financiar-bancare etc., care sunt pregătite pentru a fi analizate în vederea luării unor decizii optime;

2. *Din punct de vedere medical* - există la momentul de faţă diverse baze de date din domeniul sănătăţii (medico-farmaceutic), care au fost doar parţial analizate, cu mijloacele specifice medicinii şi care conţin un nivel mare de informaţie încă neexplorată suficient;

3. *Din punct de vedere ştiinţific-cercetare* - se cunoaşte o imensă bază de date din domeniul cercetării ştiinţifice din cele mai diverse domenii (mai exact astronomie, meteorologie, biologie, lingvistică ş.a.) care nu pot fi explorate cu mijloacele tradiţionale.

Domeniul computerelor şi al Informaticii s-a dezvoltat exponenţial având în vedere faptul că există un imens volum de date neexplorate sistematic încă, precum şi faptul că a crescut presiunea utilizării de noi metode pentru descoperirea informaţiei *ascunse* în date, informaţie care este imposibil de detectat cu mijloacele tradiţionale şi folosind doar capacitatea umană de analiză.

Datorită *tinereţii* sale, acest domeniu ştiinţific nou nu are încă o terminologie în literatura ştiinţifică străină şi românească bine definită şi acceptată de toţi. Însă există oameni care sunt interesaţi de acest domeniu nou, aceşti oameni fără pregătirea necesară sunt dornici să cunoască mai mult pentru a-şi forma o imagine cât de cât clară despre conceptul Data Mining şi, mai ales, despre posibilităţile de aplicare a acestuia.

Potrivit celor relatate, unde s-a putut, s-a utilizat şi terminologia internaţională (în limba Engleză, în original) îndeobşte folosită, pentru a uşura atât utilizarea resurselor bibliografice internaţionale precum şi căutarea pe Internet a termenilor de referinţă [58].

Având în vedere toate acestea, conceptul Data Mining nu poate fi definit ca o imagine care să ne furnizeze cât de cât fenomenul. Astfel vom încerca să redăm unele aspecte asemănătoare ale acestui concept:

- Presupune o practică a căutării automate de pattern-uri (modele, şabloane, tipare, forme etc.) în cadrul unor mulţimi mari de date, utilizând tehnici computaţionale din Statistică, învăţarea automată (machine learning) precum şi recunoaşterea formelor (pattern recognition);

- O extragere netrivială din date a informaţiei implicite, necunoscută încă, şi care poate fi folositoare;

- Reprezintă ştiinţa extragerii de informaţie folositoare din mulţimi mari de date sau baze de date;

- Presupune explorarea şi analiza unor imense cantităţi de date, prin mijloace automate sau semi-automate, în vederea descoperirii de pattern-uri folositoare;
- Procesul descoperirii automate a informaţiei - identificarea patternurilor ascunse şi relaţiilor cu datele [165], [47], [59], [60], [112].

Astfel conceptul Data Mining presupune procesul de căutare a unui *ac într-un car cu fân*, folosind un *senzor de metale*, cu scopul de a mări viteza de căutare, *automatizând* procesul respectiv.

Vom prezenta ulterior patru situaţii reale [112], ce ilustrează elocvent ce nu este Data Mining prin comparaţie cu ce ar putea fi.

- *Ce anume nu poate fi reprezentat de Data Mining?*
  - Căutarea unui anumit număr de telefon într-o carte de telefoane;
  - Căutarea unei anumite informaţii (e.g. despre bucătărie pe Google).
- *Ce anume ar putea fi reprezentat de Data Mining?*
  - Căutarea unor nume de tipul ...escu într-un stat din SUA;
  - Gruparea laolaltă a informaţiilor similare în funcţie de context (e.g. despre bucătăria franceza, italiană etc., găsite pe Google).

Pentru a înţelege mai bine acest concept un exemplu mai concludent pentru sublinierea diferenţei dintre ceea ce reprezintă uzual o căutare într-o bază de date şi conceptul Data Mining reprezintă următorul conţinut: *Cineva poate fi interesat de diferenţa între numărul de cumpărături de un anumit tip (e.g. electrocasnice) din cadrul unui supermarket în comparaţie cu un hypermarket sau, eventual, din cadrul unor supermarket-uri din două regiuni diferite*.

Astfel respectivul ia în consideraţie *a priori* ipoteza că există diferenţe între un anumit supermarket şi un hypermarket, sau între anumite vânzări dintre cele două regiuni.

În schimb în cazul Data Mining potrivit celor relatate o principală problemă poate consta de exemplu în identificarea factorilor ce influenţează volumul vânzărilor, fără a avea ca bază o ipoteză considerată a priori.

În concluzie, metodele Data Mining încearcă să identifice pattern-uri şi relaţii ascunse, care nu sunt mereu evidente. Potrivit exemplelor anterioare nu se poate pune semnul egal între o căutare/cercetare individuală a unui anumit obiect fără a se ţine cont de natura sa şi cercetarea de tip Data Mining care nu *caută* individualisme ci mulţimi de individualisme

care, într-un mod sau altul, pot fi grupate după diferite criterii. Diferenţa dintre o căutare simplă şi un anumit proces Data Mining este aceea dintre căutarea unui anumit copac şi identificarea unei anumite păduri.

Obiectivele Data Mining [42] pentru a desluşi mai clar aria sa de aplicabilitate sunt:

1. *Obiective predictive* (presupune utilizarea unei părţi din variabile pentru a prognoza una sau mai multe dintre celelalte variabile):
   - Clasificarea;
   - Regresia;
   - Detecţia deviaţiilor/anomaliilor.
2. *Obiective descriptive* (presupune identificarea de pattern-uri ce descriu datele şi care pot fi înţelese de utilizator):
   - Clustering (clusterizare);
   - Descoperirea regulilor de asociere;
   - Descoperirea de pattern-uri secvenţiale.

1. **Pentru ce utilizăm Data Mining?**

   Pentru a răspunde la o asemenea întrebare fără o prezentare prealabilă a tehnicilor Data Mining este necesar o prezentarea a două situaţii complet diferite poitrivit căreia Data Mining a fost utilizată cu succes. Pentru început se are în vedere situaţia privind rolul Data Mining în rezolvarea unei probleme fundamentale, din nefericire, a zilelor noastre [165].

   O utilizare recentă a Datei Mining se referă la următorul fapt: *Serviciile de spionaj ale Armatei SUA ar fi descoperit cu un an înainte, pe baza utilizârii tehnicilor Data Mining, indicii privind atentatele din septembrie, potrivit căreia s-ar fi identificat liderul grupului de terorişti care au pus la cale atentatul sinucigaş*. Cu părere de rău însă informaţiile nu au fost luate în seamă de autorităţi.

   Urmează o întâmplare nostimă, dar neplăcută pentru cel în cauză [59]. Pe scurt, un individ povesteşte că fiind plecat de casă, a fost sunat de compania telefonică la care era client pentru a-i aminti faptul că se presupunea că i-a fost furată cartela telefonică care a şi fost folosită într-un mod fraudulos. Compania telefonică ajunsese la această concluzie pe baza unei analize a ultimelor convorbiri telefonice efectuate utilizând

cartela respectivă, convorbiri către unele locaţii care nu se potriveau cu pattern-ul pe care compania telefonică îl construise pentru clientul său.

Acestea sunt două exemple solide pentru a lua în considere cu toată seriozitatea acest domeniu, pe cât de fascinant, pe atât de complex, al descoperirii de informaţii acolo unde cunoaştinţele umane nu mai sunt de folos.

Există un număr mare de companii care au ca obiect principal de activitate Data mining [59]. Acest lucru se datorează mai ales cererii tot mai mari de servicii de Data Mining de către piaţa financiar-economică (a se vedea, spre exemplu, domenii ca: Business intelligence (BI), Business Performance Management (BPM), Customer Relationship Management (CRM) etc.), şi piaţa de medicină (Health Informatics, e-Health etc.), fără a neglija şi alte domenii de interes precum telecomunicaţiile, meteorologia, biologia ş.a.

Pornind de la prognoza în domeniul marketingului potrivit marilor companii transnaţionale şi trecând prin analiza tendinţelor de tranzacţionare a acţiunilor de la Bursă, realizarea profilului clientului fidel, modelarea cererii de produse farmaceutice, automatizarea diagnosticării cancerului, precum şi detectarea fraudelor bancare şi a uraganelor, clasificarea stelelor şi galaxiilor ş.a, se observă o paletă diversă de domenii potrivit căreia tehnicile Data Mining sunt folosite eficient, fapt ce dă un răspuns clar întrebării: de ce se utilizează Data Mining?

Însă nu trebuie considerat faptul că Data Mining poate rezolva orice problemă de descoperire de informaţie utilă în date. Identic ca şi mineritul originar, şi în cazul Data Mining este posibil să se *sape* în *muntele* de date chiar daca nu se dă peste filonul de aur al cunoaşterii. Descoperirea de cunoştinţe sau de informaţii utile depinde foarte mult de anumiţi factori, începând cu *muntele* de date în care se *minereşte* şi finalizând cu *uneltele* de Data Mining utilizate şi, desigur, să nu uităm priceperea *minerului*. Dacă nu există *aur* în munte, degeaba se sapă. Filonul de aur, dacă există, este necesar să fie identificat şi evaluat corect şi concret, iar apoi, dacă este eficientă exploatarea sa, este necesar să fie executată cu unelte de minerit corespunzătoare.

2. **Cum se *minereşte* în date?**

   Procesul de *minerit* în date, are la bază trei etape definitorii ai procesului de Data Mining:

   - Explorarea datelor, presupune *curăţarea* şi transformarea datelor, selectarea de sub mulţimi de date şi a caracteristicilor, acolo unde avem un mare număr de

variabile ş.a.;

- Construirea modelului şi validarea acestuia face referire la considerarea diferitelor modele şi alegerea aceluia cu cea mai bună performanţă în prognoză - evaluarea competitivă a modelelor (conform literaturii străine denumit *competitive evaluation of models*);
- Aplicarea modelului la date noi, în vederea producerii de prognoze sau estimaţii corecte pentru problemele cercetate.

Conform specialistului R. Groth [59], se pot identifica cinci etape ale *mineritului* în date:

(a) **Pregătirea datelor (data pre-processing).** Înainte de utilizarea unei tehnici de Data Mining, este absolut necesară pregătirea datelor *brute* (raw data) în vederea analizării eficiente. Se cunosc mai multe trăsături ale pregătirii iniţiale a datelor înaintea procesării propriu-zise cu ajutorul tehnicilor Data Mining [112]. Se are în vedere pentru început problema calităţii datelor (şi anume existenţa zgomotului (noise), a valorilor extreme sau aberante (outlier/anomaly), a valorilor lipsă (missing values), a datelor duplicat, introduse incorect sau expirate ş.a.).

În concordanţă cu problemele detectate privind calitatea datelor, se are în vedere rezolvarea acestora cu metode specifice. Luând ca exemplu existenţa *zgomotului*, mai exact acele distorsiuni ale valorilor (măsurătorilor) originale produse de perturbări aleatoare, se utilizează diferite tehnici de *filtrare*, care au ca scop îndepărtarea sau reducerea efectului distorsiunilor.

Un alt exemplu este cel al procesării semnalelor (signal processing), în afara filtrelor electronice (hard), filtrele *matematice* (soft), adică algoritmi matematici care sunt folosiţi în vederea modificării componentei armonice a semnalului. În concluzie, dacă există unele valori extreme, adică cele care se abat foarte mult de la media celorlalte valori ale datelor, se are în vedere îndepărtarea acestora, sau utilizarea unor parametri (statistici) care să nu fie aşa de sensibili la aceste valori extreme (utilizarea medianei în locul mediei, care este *sensibilă* la valorile extreme).

Cazul valorilor lipsă este foarte des întâlnit în practica Data Mining şi are o multitudine de cauze. Astfel, se utilizează diferite metode, ca: eliminarea obiectelor sau instanţelor care au atribute lipsă, estimarea valorilor lipsă şi înlocuirea