UNIUNEA EUROPEANĂ

GUVERNUL ROMÂNIEI

Instrumente Structurale
2007-2013

Platformă de e-learning și curriculă e-content pentru învățământul superior tehnic

Proiect nr. 154/323 cod SMIS – 4428 cofinanțat de prin Fondul European de Dezvoltare Regională “Investiții pentru viitorul dumneavoastră”.

Programul Operaţional Sectorial Creşterea Competitivităţii Economice - POS CCE

UNIUNEA EUROPEANĂ

GUVERNUL ROMÂNIEI

Instrumente Structurale
2007-2013

Platformă de e-learning și curriculă e-content pentru învățământul superior tehnic

Baze de date

20. Sintetizarea şi gruparea datelor

Introducere

Pentru prelucrarea datelor este necesară sintetizarea şi gruparea lor. Sintetizarea datelor se obține în urma centralizării, pentru a permite prelucrări statistice. Prelucrarea cuprinde operații de rafinare sau transformare, cu ajutorul cărora se realizează trecerea de la datele individuale la indicatori sintetici. Gruparea permite organizarea și aranjarea înregistrărilor după grup, de exemplu vânzarile pe zone geografice sau tipuri de produse. Grupurile pot fi pe mai multe niveluri, astfel încât să se poată identifica ușor relațiile dintre grupuri și să se poată găsi ușor informațiile dorite. De asemenea, gruparea se poate utiliza pentru a calcula informații de sinteză (totaluri, procente, functii statistice, etc.). Scopul final este obținerea de rapoarte specifice pentru adoptarea deciziilor. Rapoartele simple se pot obține prin interogări pe un singur tabel, dar de cele mai multe ori rapoartele se fac prin interogări complexe, pe mai multe tabele, care stochează date corelate între ele.

Analiza şi interpretarea datelor

- **Analiza datelor** are ca scop descoperirea relațiilor dintre datele sintetizate: tipare, asocieri, corelații pe plan structural, funcțional şi cauzal. Cea mai simplă formă de analiză a datelor este compararea datelor sintetizate cu date similare.
- **Interpretarea datelor** produce cunoştințe noi, care se vor adăuga celor existente. Un sistem informatic suport, care să susțină procesele de transformare a datelor în informații şi a informațiilor în cunoştințe, va trebui să aibă drept principale funcționalități: gestionare de date, gestionare de modele, gestionare de cunoştințe, gestionare de interacțiuni, dintre utilizator şi sistem, pe de o parte şi între date, modele şi cunoştințe, pe de altă parte. Interpretarea datelor este un proces în care se face apel la cunoştințele cu caracter general, fundamental şi specific asociate domeniului respectiv precum şi la experiența în domeniu.

Tehnici de data mining

- Tehnicile de observare analitică a datelor folosesc o tehnologie modernă numită *data mining*. Procesul de observare analitică permite obținerea unor tipare, corelații şi chiar modele, din care se pot deduce tendințe, se poate specifica, cu o anumită probabilitate, evolutia fenomenelor în perioada următoare, permițând interpretarea datelor.Interpretarea datelor este un proces cognitiv, care conduce la o apreciere generală a situației, la identificarea unor probleme sau sesizarea unor oportunități.
- Tehnicile de *data mining* au fost grupate în trei categorii, în funcție de tipul de probleme pe care le pot modela:
- *Clasificarea şi regresia,* constând în construirea de modele pentru previzionarea apartenentei la un set de clase (clasificare), sau a unor valori (regresie).
- *Analiza asocierilor şi succesiunilor*, această tehnică generează modele descriptive, care evidențiază reguli de corelatie între atributele unui set de date.
- *Analiza de tip cluster* este o tehnică descriptivă, utilizată pentru gruparea entităților similare, dintr-un set de date, sau pentru evidențierea entităților care prezintă diferențieri substanțiale față de un grup.

Interogari SQL folosind funcții de grup

Funcțiile de grup returnează rezultate bazate pe grupuri de înregistrări, nu pe o singură înregistrare. Gruparea se face folosind clauza *GROUP BY* într-o comandă *SELECT* şi în acest caz toate elementele listei trebuie cuprinse în clauza de grupare. Funcțiile de grup pot fi apelate şi în clauza *HAVING*, dar nu pot fi apelate în clauza *WHERE*. Se poate folosi operatorul *DISTINCT* pentru a sorta numai elementele distincte din listă, dar se poate folosi şi operatorul *ALL* pentru a considera şi înregistrările duplicate.

Tipuri de funcții de grup

- **COUNT(* | [DISTINCT/ALL] expr)** – returnează numărul de înregistrări întoarse de interogare;

  Dacă se foseşte * , se numără toate înregistrările, inclusiv cele nule, iar dacă se foloseşte **expr** se numără numai înregistrările **not null**.
- **AVG([DISTINCT/ALL] expr)** – returnează valoarea medie a lui **expr**, ignorând valorile nule;
- **MAX([DISTINCT/ALL] expr)** – returnează valoarea maximă pentru **expr;**
- **MIN([DISTINCT/ALL] expr) –** returnează valoarea minimă pentru **expr;**
- **SUM([DISTINCT/ALL] expr) –** returnează suma valorilor pentru **expr;**
- **VARIANCE([DISTINCT/ALL] expr)** – returnează variația standard pentru **expr,** ignorând valorile nule;
- **STDDEV([DISTINCT/ALL] expr) –** returnează deviația standard pentru **expr** ignorând valorile nule.

Exemple:

- SQL> SELECT id_dep,count(*),count(comision), count(all comision), count(distinct comision) FROM angajati GROUP BY id_dep;
- SQL> SELECT id_dep, avg(salariu), avg(all salariu), avg(distinct salariu) FROM angajati GROUP BY id_dep;
- SQL> SELECT a.id_dep,b.den_dep,max(salariu) FROM angajati a,departamente b WHERE a.id_dep=b.id_dep GROUP BY a.id_dep,b.den_dep;
- SQL>SELECT id_dep, min(salariu + nvl(comision,0)) venit_minim FROM angajati GROUP BY id_dep;
- SQL> SELECT id_dep,sum(salariu),sum(distinct salariu), sum(comision) FROM angajati GROUP BY id_dep;
- SQL> SELECT id_dep,variance(salariu),variance(comision) FROM angajati GROUP BY id_dep;
- SQL> SELECT id_dep,stddev(salariu), stddev(distinct salariu), stddev(comision) FROM angajati GROUP BY id_dep;