Modelul semistructurat al datelor

Asist. Adriana REVEIU
Catedra de Informatica Economica, A.S.E. Bucuresti

The necessity of data integration for various data management models (relational, object-oriented) and still for data with the same data management model but with different schemas, was a request for the development of the semi structured data model. More and more applications demand the using of data that cannot be restricted to any classical data management model. The semi structured data model gives flexibility, increase the efficiency of querying information, assists data integration based on various data management models.

Keywords: *data integration, semi structured data model, OEM (*Object Exchange Model*), UnQL* (*Unstructured Query Language*), *Lorel.*

Generalitati

Modelul relational si cel orientat-obiect sunt folosite de mult tip pentru modelarea datelor nu neaparat pentru ca sunt cele mai naturale solutii, ci pentru ca au în spate un formalism bine definit. În timp însa, datorita cerintelor de integrare a datelor descrise folosind modele diferite, dar mai ales datorita dezvoltarii Internetului, modelele traditionale au devenit insuficiente impunându-se ca acuta folosirea unui model care sa lege cele doua lumi distincte: relationala si orientata-obiect. Termenul utilizat în legatura cu acesta este modelul semistructurat al datelor. Mai mult, aplicatiile Internet realizeaza operatii de gestiune a datelor care nu sunt comune si aplicatiilor care folosesc baze de date traditionale cum ar fi: transformarea datelor, interpretarea *query*-urilor, transportul datelor si prelucrarea bazata pe fluxuri. Datele semistructurate au devenit de curând un important subiect de studiu din mai multe motive.

În primul rând, exista surse de date pe care am vrea sa le tratam ca baze de date dar carora nu le putem impune constrângeri având la baza o schema, ca în cazul bazelor de date traditionale. Chiar si datele structurate au structura care diferita de la un caz la altul sau structura modificabila în timp. Dezvoltarile din domeniul multimediei au impus de asemenea folosirea de noi tipuri de date în domeniul tehnologiei bazelor de date conventionale. Unele dintre acestea necesita doar extensii ale modelelor de date existente, deoarece ele impun optimizarea tehnicilor de manipulare si interogare a noilor tipuri de date, dar altele nu pot fi încadrate în modelele clasice de gestiune a datelor. Cel mai evident exemplu de date care nu pot fi încadrate într-o schema sunt datele manipulate prin intermediul Web-ului. Majoritatea *query*-urilor Web exploateaza tehnicile de regasire a informatiilor pentru a determina paginile individuale care au continut ce corespunde criteriului solicitat. Dar numai o mica parte din resursele Web sunt structurate astfel încât sa permita rezolvarea *query*-urilor, Web-ul nefiind conform nici unui model standard. Astfel ca s-a simtit nevoia de a gasi o metoda pentru descrierea structurii Web-ului în vederea rezolvarii acestei probleme.

Al doilea motiv este dat de necesitatea schimbului de informatii între aplicatii care folosesc formate diferite de stocare si gestiune a informatiilor necesitând astfel transformarea datelor. Rationamentul este urmatorul: nici unul din modelele de date existente nu este acceptabil din toate punctele de vedere si în anumite situatii este dificil sa se converteasca datele dintr-un model în altul. În plus, datele în format electronic se afla în medii diferite interconectate care trebuie sa comunice între ele. Dar si când se lucreaza cu date structurate poate fi util ca ele sa fie privite ca date semistructurate, din motive care tin de manipularea acestora, fiind util sa existe un format flexibil pentru schimbul datelor între diferite tipuri de baze de date. O parte din aceste informatii se regasesc sub forma datelor nestructurate, ca de exemplu sunetele, imaginile, secventele video si chiar unele documente text, iar alta parte se regasesc sub forma informatiilor structurate memorate în baze de date relationale sau orientate obiect.

Acest fapt a facut ca modelele relational si obiectual sa nu mai fie suficiente pentru organizarea datelor în mediul Internet si nu numai. În general un utilizator nu poate scrie un *query* pentru o baza de date fara a-i cunoaste schema de organizare a datelor. Schemele de organizare sunt netransparente pentru utilizatori si rationamentul folosit la proiectarea lor este adesea greu de determinat. Poate fi util în rezolvarea problemei sa se foloseasca o varianta de interogare a datelor fara a cunoaste în totalitate schema bazei de date. Au fost dezvoltate limbaje care permit interogarea simultana a datelor si a schemelor de organizare a lor, în contextul sistemelor bazelor de date orientate-obiect si relationale, dar aceste limbaje nu au flexibilitatea sa manipuleze constrângeri complexe.

Conceptul de date semistructurate

În general, prin termenul de date semistructurate sunt denumite date care nu respecta formate stricte cum ar fi cele impuse de modelele bazelor de date relationale sau orientate obiect. În mod evident o astfel de definitie este imprecisa. Datele sunt semistructurate nu au o *structura rigida* (de exemplu datele Web), sunt *combinate* din câteva surse eterogene (ca în cazul *data warehousing*), nu au o *structura implicita* asociata datelor sau au o *structura partial specificata*.

Modelele orientat-obiect si relational au o schema fixa pentru fiecare clasa sau fiecare relatie. Datele semistructurate permit o *flexibilitate sporita*, fiind un amestec al celor doua concepte: clasa si relatia. Datele semistructurate poarta informatii despre schema lor si aceasta schema poate varia în timp, în cadrul unei singure baze de date. Modelul care nu se bazeaza pe nici o schema are un rol special în sistemele bazelor de date.

Datele în acest model sunt *autoreferite* însemnând ca schema este atasata datelor. Bazele de date cu scheme identice pot fi combinate cu usurinta dar în cazul în care o baza de date este relationala si alta orientata-obiect apare problema mostenirii bazei de date. O posibila solutie este conectarea celor doua baze de date mostenite printr-o interfata, interfata folosind date semistructurate si utilizatorul putând realiza interogarea prin intermediul unui limbaj de interogare. Bazele de date sunt construite translatând datele de la sursa în date semistructurate. În modelul datelor semistructurate, informatiile care sunt în mod normal asociate unei scheme sunt incluse în interiorul datelor. În aceste baze de date nu exista o distinctie clara între date si schema, iar gradul de structurare depinde de aplicatie. În anumite forme ale modelului semistructurat exista o schema distincta, în altele nu.

Avantajele aduse de modelul semistructurat al datelor sunt urmatoarele:

? *Flexiblitatea* simplifica integrarea bazelor de date cu date similare dar având scheme de organizare diferite. Spre deosebire de alte modele, care au o schema de descriere a datelor, modelul semistructurat al datelor este fara schema justificând astfel flexibilitatea lui. Flexibilitatea este utila în vederea integrarii informatiilor, mai ales în cazul problemelor legate de mostenirea bazei de date. Aceasta situatie se întâlneste, de exemplu la fuzionarea a doua companii care au datele stocate în baze de date cu structuri diferite.

? *Îmbunatatirea eficientei regasirii informatiilor (pe Web)*. Documentele Web si cele pastrate în biblioteci digitale sunt semistructurate. Spre deosebire de fluxul nestructurat de date (întâlnit la imagini, sunete, video), datele semistructurate au o structura si spre deosebire de datele structurate (bazele de date relationale sau orientate-obiect), datele semistructurate nu au o schema absoluta sau o clasa fixa, fiecare obiect continând propria "schema". Iregularitatea structurala nu implica inexistenta similaritatilor structurale între obiecte. Din contra, în mod obisnuit, obiectele semistructurate care descriu acelasi tip de informatii au structura similara.

? *Posibilitatea integrarii datelor*. Modelul semistructurat al datelor este un model pentru integrarea bazelor de date si este folosit pentru descrierea datelor aflate în doua sau mai multe baze de date care contin date similare în diferite scheme de organizare.

Modelarea datelor semistructurate

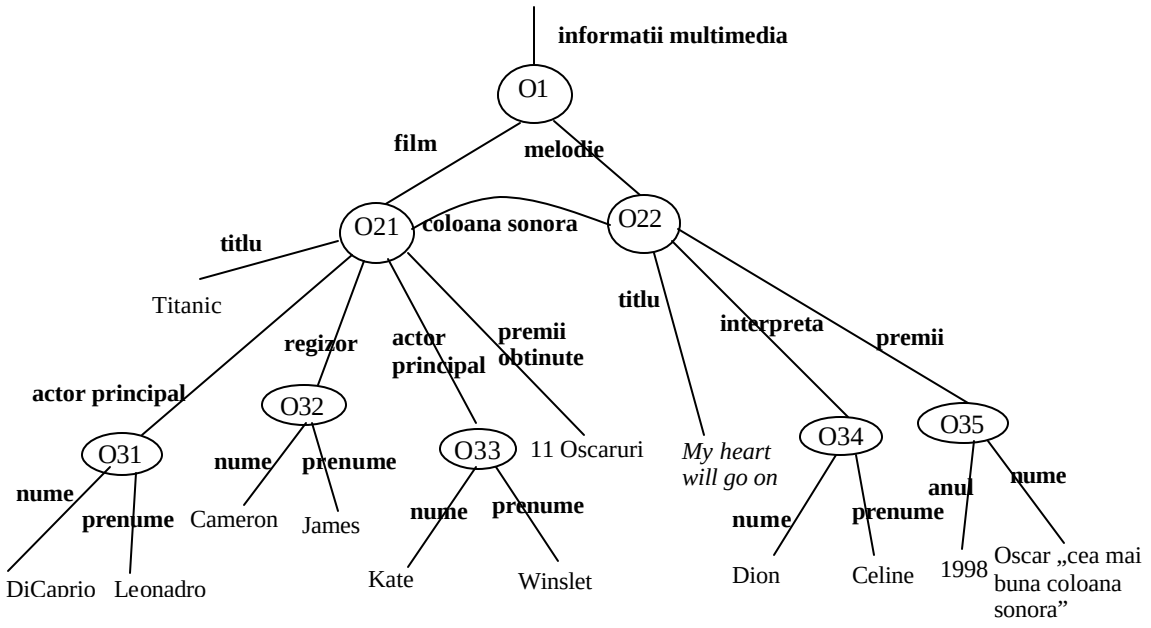
Datele semistructurate sunt modelate în general sub forma structurilor de tip graf sau arbore

etichetat cu nodurile reprezentând obiecte iar arcele reprezentând atribute. În plus, arcele modeleaza natural relatiile de subordonare dintre doua obiecte. Modelul de facto pentru datele semistructurate este OEM (*Object Exchange Model*), model propus în proiectul pentru integrarea datelor TSIMMIS. OEM este autodescriptibil, nefiind necesara definirea anterioara a structurii unui obiect si nici nu exista o schema fixa de reprezentare a datelor.

Fiecare obiect din OEM poate avea un identificator (*id*), o eticheta, un tip si o valoare. *Id*-urile obiectelor pot fi simboluri cu sau fara înțelesuri speciale. De exemplu, daca un obiect este o pagina Web, URL-ul paginii poate fi folosit ca *id*-ul obiectului. *Etichetele nodurilor* sunt siruri care expliciteaza rolul obiectului în aplicatii. Etichete joaca aici doua roluri: identifica un obiect si identifica sensul obiectului. Obiectele pot fi *atomice* sau *complexe* (seturi de obiecte). Valorile obiectelor atomice au tipuri atomice (valori întregi, reale, siruri de caractere, imagine, sunet, etc.), iar cele complexe au ca valori seturi de obiecte, reprezentate prin perechi de tipul atribut-obiect. Astfel ca un obiect complex va avea o definire recursiva deoarece valoarea unui obiect este parte a sa. Nodurile frunza au asociate întotdeauna valori atomice.

Etichetele arcelor reprezinta *atribute*, sunt descrise prin siruri de caractere si reprezinta un set de proprietati descriptive. O proprietate poate fi folosita într-un arc pentru a descrie nodurile învecinate.

Exemplificarea modelarii unei parti dintr-o baza de date multimedia folosind OEM este prezentata în figura 1.



unde: **oij** sunt *id*-urile nodurilor grafului, fara o anumita semnificatie;

.

Fig. 1 Modelarea unei baze de date multimedia

```
                                Informatii: &o1
{film: &o21
      {titlu: &o30 “Titanic”,
       actor principal: &o31
           {nume: &o41 “DiCaprio”,
      prenume: &o42 “Leonadro”},
        regizor: &o32
      {nume: &44 “Cameron”,
             prenume: &o45 “James”},
```

```
       actor principal: &o33
     {nume: &o46 "Kate",
      prenume: &o47 "Winslet"},
        premii obtinute: &o48 "11 Oscaruri"},
melodie: &o22
     {titlu: &49 "My heart will go on",
      interpreta: &o34
          {nume: &o50 "Dion",
      prenume: &o51 "Celine"},
       premii: & o35
     {anul: &o52 "1998",
       nume: "Oscar pentru cea mai buna coloana sonora"},
        coloana sonora: &o22}}
```

Scheme de organizare a datelor semistructurate

Flexibilitatea în descrierea informatiilor adusa de modelarea datele semistructurate are si dezavantaje si anume dificultatea formularii unei interogari relevante. S-a încercat asocierea unor scheme în modelarea datelor semistructurate, în acest sens existând doua abordari:

? *scheme flexibile* – descriu aprioric datele; schemele flexibile fiind concepute astfel încât sa permita flexibilitatea în descrierea datelor; se foloseste tipul *void* pentru manipularea oricarui tip de data;

? *scheme rigide* – descriu datele cu exactitate folosite pentru analiza datelor; schema este recalculata ori de câte ori se produce o modificare a informatiei memorate.

Limbaje de interogare pentru datelor semistructurate

Principalele elemente caracteristice ale unui limbaj de interogare pentru datele semistructurate sunt:

? *putere de expresie*: pentru reprezentarea datelor relationale ca date semistructurate, limbajul semistructurat trebuie sa acopere operatiile corespunzatoare unui limbaj relational standard, dar în plus trebuie sa dispuna de facilitati de reorganizare a datelor, astfel încât aceeasi informatie sa poata fi regasita sub o alta structura;

? *semantica*: cererile trebuie optimizate astfel încât sa poata tine cont de semantica din reprezentarea sintactica a datelor fiind necesara o semantica precisa pentru transformarea si optimizarea cererilor;

? *schematizarea* : limbajul trebuie sa poata recunoaste structurile definite pentru a le putea manipula;

? *compunerea*: datele obtinute în urma unei interogari pot fi folosite ca date de intrare în alte interogari, motiv pentru care limbajele trebuie sa fie transparente din punctul de vedere al referintelor.

Au fost propuse, într-o serie de proiecte de cercetare pe aceasta tema câteva limbaje de interogare pentru datele semistructurate. Doua dintre acestea sunt limbajul orientat-obiect *Lorel*, derivat din OQL si limbajul *UnQL* (*Unstructured Query Language*). Domeniul datelor semistructurate este unul în studiu si va avea un impact mare asupra unei multitudini de aplicatii, dar mai ales asupra aplicatiilor multimedia si a celor pe Internet.

Bibliografie

📖 Abiteboul S., Buneman P., Suciu D. - *Data on the Web: From Relations to Semi structured Data and XML* - Morgan Kaufmann Publishers, San Francisco, California, 1999;

📖 Abiteboul S., Benjelloun O., Milo T. - *Web services and data integration* – International Conference on Web Information Systems Engineering, 2002;

📖 Buneman P. – *Semi structured Data* – în SIGMOD/PODS'97, Mai 1997;

📖 Nwosu K., Thurai-Singham B., Berra P.B. – *Multimedia database systems – Design and implementation Strategies* – Kluwer Academic Publishers, Massachusetts, USA, 1996.