Modelul vectorial de organizare a colectiilor de date
(Partea I)

Prof.dr. Ion LUNGU, lect. Ileana TANASE
Catedra de Informatica Economica, A.S.E. Bucuresti

Modelul vectorial este unul dintre cele mai fundamentate modele privind organizarea colectiilor de date si este implementat în sistemul experimental SMART. Ideea esentiala este aceea ca termenii de indexare sunt priviti drept coordonatele unui spatiu de informatie multidimensional.
Cuvinte cheie: *entitate, interogare, similitudine.*

Notiuni generale

Modelul vectorial este cel care a avut cea mai mare influenta în dezvoltarea teoriei de regasire a informatiilor (IR). Este cel care a asigurat baza pentru o serie de experimente realizate de Salton si de colaboratorii sai si care s-a finalizat prin implementarea în sistemul SMART. Enti-tatile si interogarile sunt reprezentate de vectori, fiecare element al acestora fiind dat de un numar precis ce reprezinta va-loarea unui termen index asociat la entitate.

Multimea completa a valorilor dintr-un vector descrie pozitia entitatii sau intero-garii în spatiu. Între entitati si interogari se pot stabili similitudini bazate pe compa-rarea vectorilor atasati. Ca o masura de similitudine se poate considera atât produ-sul intern a doi vectori, cât si cosinusul dintre acestia. Aceasta interpretare geome-trica a cautarii este destul de intuitiva si reprezinta baza pentru operatiile de cautare a entitatilor. Modelul vectorial se concen-treaza pe potrivirea dintre interogari si enti-tati, iar cautarea entitatilor nu tine seama de relatiile care exista între acestea. Ori-cum, este usor de aratat ca entitatile similare vor fi raspuns la aceleasi cereri si vor fi cautate împreuna. Aceasta observatie furnizeaza baza ipotezei clusterilor de van Rijbergen si Sparck Jones (1973) care au sugerat potrivirea unei clase de entitati pentru a obtine o eficacitate de cautare mai mare. Teoria clusterilor se bazeaza pe relatia de similitudine dintre entitati si a fost elaborata mult mai pe larg de Willett în 1988.

Configurarea spatiului de entitati

Regasirea elementelor reprezinta caracte-ristica principala a organizarii datelor si în consecinta gasirea unui sistem optim de indexare este foarte necesara. Matematic se incearca obtinerea unui vocabular de inde-xare cât mai bun, prin calculul densitatii spatiului S. Consideram entitatile E_i, fieca-re identificata prin unul sau mai multi ter-meni index, notati $T_j, i,j \in N^*$.

Reprezentarea tridimensionala a spatiului termenilor index va arata ca în figura 1, în conditiile în care fiecare element al spa-tiului este identificat prin 3 termeni index distincti. Daca entitatile admit *t* termeni index, atunci exemplul va fi extins la *t* dimensiuni, când *t* termeni index diferiti sunt prezenti. În acest caz fiecare entitate E_i va fi reprezentata printr-un vector de dimensiune t.

$$E_i=(e_{i1},e_{i2},...,e_{it}), \qquad (1),$$

unde e_{ij} reprezinta valoarea termenilor in-dex T_j.

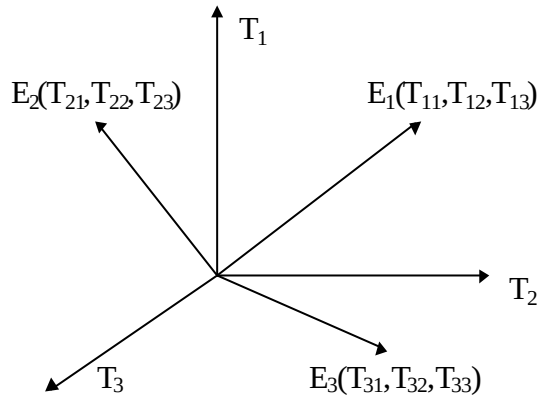

Fig. 1. Reprezentarea vectorilor în spatiul entitatilor

Analog interogarile ar putea fi si ele reprezentate într-o forma vectoriala, sau intr-o forma de declaratii booleene. Astfel o interogare Q ar putea fi formulata astfel:

$Q=(q_a,q_b,...,q_r)$ (2) sau

$Q=(q_a$ and $q_b)$ or $(q_c$ and $q_d$and...) or... (3)

unde q_k reprezinta termenii asociati intero-garii Q. O reprezentare formala a vectorilor din ecuatiile (1), (2) si (3) este obtinuta incluzând în fiecare vector pentru toate componentele o valoare care sa creeze distinctii între termeni. Daca w_{ek} si w_{qk} reprezinta valoarea termenului t_k în entitatea E sau în interogarea Q, atunci vectorii asociati pot fi scrisi sub forma $E=(e_1:w_{e1}; e_2:w_{e2}; ...; e_t:w_{et})$ si $Q=(q_1:w_{q1}; q_2:w_{q2}; ...; q_t:w_{qt})$ (4).

Vom presupune ca w_{ek} (sau w_{qk}) sunt egale cu 0, daca termenul k nu este asignat la entitatea E (sau interogarea Q) si w_{ek} (sau w_{qk}) sunt egale cu 1 pentru termenii asignati.

Coeficientul de similitudine

Pentru doua entitati cu vectorii index asociati se poate calcula un coeficient de similitudine dintre ele, $s(E_i,E_j)$ care va reflecta gradul de asemanare a termenilor corespondenti. Ca o masura de similitudine $s(E_i,E_j)$ ar putea fi produsul intern al celor doi vectori $< E_i, E_j >= \sum_{k=1}^{t} w_{ik} \cdot w_{jk}$ sau al-ternativ unghiul dintre doi vectori

$$\cos(E_i, E_j) = \frac{< E_i, E_j >}{IIE_i II \cdot IIE_j II}.$$

Daca termenii asociati celor doi vectori sunt identici, unghiul va fi zero, iar simili-tudinea maxima.

Analog se poate calcula si o valoare de similitudine interogare-entitate obtinuta tot ca produsul intern al celor doi vectori, sau ca o valoare de cosinus a vectorilor. $s(Q,E) = \sum_{k=1}^{t} w_{qk} \cdot w_{ek}$. Daca valorile termenilor index sunt restrictionate la 0 si 1, asa cum am s-a sugerat anterior produsul intern a doi vectori masoara numarul de termeni care sunt împreuna asignati la ambii vectori în acelasi timp.

În practica, s-a dovedit util sa se asigure un grad mai mare de discriminare între termenii asignati pentru reprezentarea continutului. De aceea sunt permise nu numai valori de 0 si 1 pentru valorile termenilor, ci valori care variaza între 0 si 1. Asignarea unor valori apropiate de 1 este folosita pentru termenii cei mai importanti, iar valori apropiate de 0 sunt folosite pentru termenii cu o mai mica importanta. În aceste circumstante ar putea fi destul de util daca s-ar folosi vectori normalizati. Factorul de normalizare ar putea fi $\frac{w_{ek}}{\sqrt{\sum_{vector}(w_{ei})^2}}$ pentru entitati, si $\frac{w_{qk}}{\sqrt{\sum_{vector}(w_{qi})^2}}$ pentru interogari.

Desi metoda vectoriala a fost cea de la care s-a pornit, în cazul vectorilor cu multe dimensiuni reprezentarea lor completa cu originea în O este imposibil de realizat. S-a renuntat la aceasta idee, distanta relativa dintre vectori fiind conservata prin normalizarea tuturor lungimii vectorilor la 1 si considerarea proiectiei vectorilor pe înfasuratoarea spatiului reprezentata prin sfera unitate. Acceptarea normalizarii vectorilor conduce la o relatie de similitudine de forma

$$s(E_1, E_2) = \frac{\sum_{k=1}^{t} w_{ek}^1 \cdot w_{ek}^2}{\sqrt{\sum_{k=1}^{t} (w_{ek}^1)^2 \cdot \sum_{k=1}^{t} (w_{ek}^2)^2}}$$ pentru entitati si

anolog pentru s(Q,E), unde Q este o interogare. Aceste formule au fost folosite în mod extins în sistemul de regasire experimental SMART.

În acest caz, fiecare entitate poate fi înfa-tisata printr-un singur punct a carui pozitie este data de intersectia dintre vectorul unitate si înfasuratoarea spatiului. Doua entitati cu termenii index asociati similari vor fi reprezentati prin puncte aflate la distante mici unul fata de altul. În general, distanta dintre doua puncte entitate în spatiu este invers corelata cu similitudinea dintre vectorii corespondenti. (Similitudine mare implica distanta mica între puncte si invers).

Problema ar putea fi privita si invers: decât sa reprezentam optim entitatile într-un spatiu dat, mai bine am corecta un spatiu astfel încât sa devina optim pentru entita-tile date. Daca nu se cunoaste nimic special despre elementele colectiei s-ar putea pre-supune ca spatiul optim de entitati este unul unde entitatile sunt grupate daca furnizeaza acelasi raspuns la o interogare, asigurând ca vor fi încarcate împreuna. În mod contrar, elementele singulare care nu au similitudini cu alte elemente sa apara bine separate în spatiul de entitati. O astfel de situatie este aratata în figura 2, unde distanta dintre doua reprezentari x a doua entitati este invers legata de similaritatea dintre vectorii index corespunzatori.

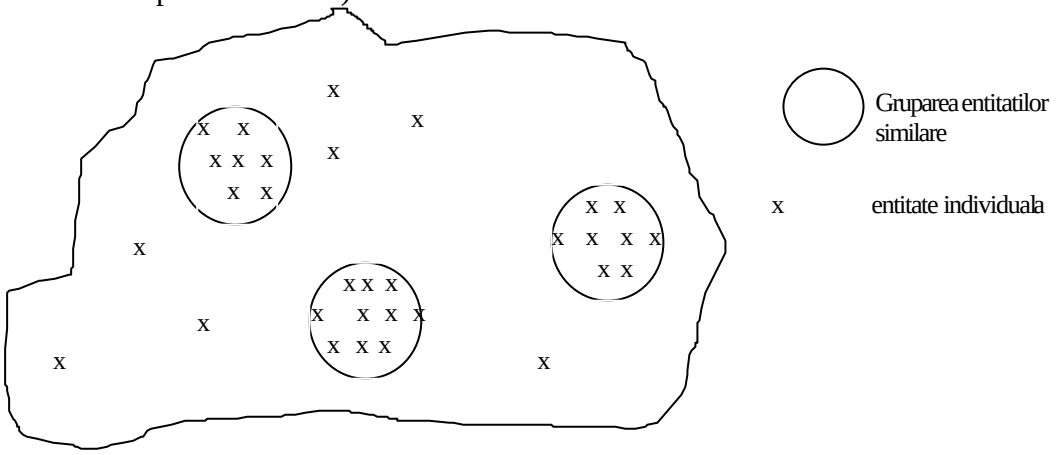

Fig. 2. Spatiul ideal de entitati

Desi configurarea entitatilor asa cum se arata în figura 2 poate fi intr-adevar cea optima, în practica nu exista o cale de realizare actuala a unui astfel de spatiu, deoarece în timpul procesului de indexare este dificil de anticipat care va furniza în timp aproape acelasi lucru. Deoarece acest spatiu optim este foarte greu de obtinut, s-a încercat o separare a tuturor entitatilor existente în spatiu asa cum se arata în figura 3. Acest lucru este obtinut daca similitudinea dintre entitati este mica. În acest sens s-a considerat functia

$$F = \sum_{i=1}^{n} \sum_{j=1}^{n} s(E_i \cdot E_j) \qquad (1),$$

care se doreste a fi minimizata. $s(E_i,E_j)$ este similitudinea dintre entitatile E_i si E_j.

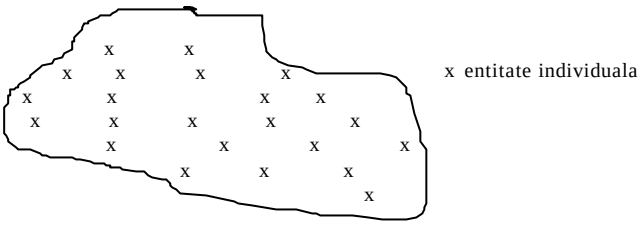

Fig. 3. Spatiul cu separare maxima între perechi de entitati

Se poate observa ca atunci când functia din exemplul (1) este minimizata, media similitudinilor dintre perechile de entitati este cea mai mica, astfel garantând ca fiecare entitate poate fi obtinuta când s-a localizat suficient de aproape de o interogare utili-zator "fara a fi nevoie sa îi încarcam si vecinii". Aceasta asigura o înalta precizie de cautare si da posibilitatea încarcarii unei entitati relevante fara sa încarcam un numar de entitati nerelevante din vecina-tatea sa. Aceasta reprezentare este foarte buna atunci când interogarea are ca rezultat o entitate, sau entitati aflate în aceeasi arie. În cazul în care interogarea este formata din mai multe entitati relevante dispersate în spatiu, încarcarea lor implica încarcarea unui numar mare de entitati nerelevante. Pe de alta parte, nici functia din exemplul (1) nu este prea usor de calculat, tinând cont ca sunt necesare n^2 comparari, pentru o colectie de n entitati. Din aceasta ratiune, un spatiu de entitati care sa grupeze entita-tile în clase este demn de luat în conside-rare. Fiecare clasa existenta este prezenta si cu un centru propriu.

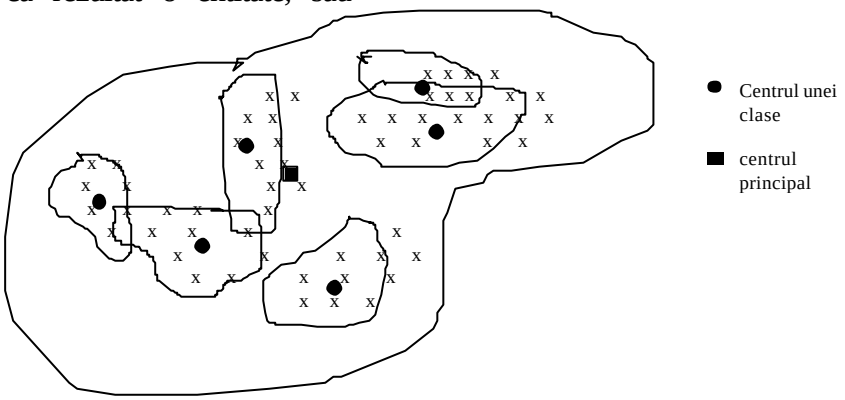

Fig. 4. Clasele cu centre

Un astfel de spatiu de entitati separate în clase a fost aratat în figura 4, unde entita-tile care apartin unei clase sunt grupate , iar centrul fiecarei clase a fost reprezentat prin puncte negre.

Pentru o clasa de entitati K care comprima m entitati, centrul clasei K poate fi definit ca fiind media aritmetica a componentelor tuturor entitatilor (reprezentate vectorial).

$$c_j = \frac{1}{m}\left(\sum_{i=1}^{m} w_{i1}, \sum_{i=1}^{m} w_{i2}, \ldots, \sum_{i=1}^{m} w_{it}\right) \qquad (2)$$

unde cj este o reprezentare vectoriala pe componenete, iar m este numarul entitati-lor din clasa j. Corespunzator centrului fie-carei clase se poate defini si un centru principal, ca fiind "centrul centrelor". A-cest centru principal poate fi obtinut din centrele claselor în aceeasi maniera în care centrele claselor sunt calculate pentru enti-tatile individuale. Deci centrul principal va fi pur si simplu media aritmetica a centre-lor claselor. Notatia folosita pentru centrul principal este c^*.

Într-un spatiu entitate de clase, masura densitatii spatiului compusa din suma tutu-ror similaritatilor entitatilor pereche (ecua-tia (1)), poate fi înlocuita ca suma tuturor coeficientilor

de similitudine dintre fiecare entitate si centrul principal:

$Q = \sum_{i=1}^{n} s(c^*, E_i)$, unde c^* este centrul principal, iar n este numarul total de entitati. Daca pentru functia din exemplul (1) erau necesare n^2 comparari, evaluarea functiei Q se poate face doar cu n comparari.

Concluzii

Gruparea entitatilor în clase trebuie facuta dupa criterii precise. Obtinerea unui tip de grupare asa cum se arata în spatiul separat din fig(3) contribuie la o buna performanta de regasire. Daca tinem seama ca entitatile aflate în apropierea unui aceluiasi centru de clasa arata caracteristici de relevanta identice (respectând majoritatea interogari-lor utilizator) atunci cea mai buna perfor-manta de regasire ar trebui sa fie obtinuta intr-un spatiu de clase individuale compac-te, dar cu distante intercentre mari. Aceasta se poate realiza daca:

(a) similitudinea dintre perechile de entitati dintr-o clasa este mare

(b) similitudinea medie dintre centrele claselor este micsorata.

Observatie: clasele sunt separate clar unele de altele, iar într-o clasa entitatile sunt a-propiate

Bibliografie

1. Salton G., Automatic Information Organization and Retrieval, McGraw-Hill, New York, 1975

2. Salton G., Yang C.S., Contribution to the theory of indexing,American Elsevier, New York,1980

3. Wong A., An investigation of the effects of differnt indexing methods on the document space configuration, Cambridge, England,1987

4. Van Rijsbergen C.J., Information retrieval, Butterworths,1988