

Modelul vectorial de organizare a datelor

Prof.dr. Ion LUNGU

Catedra de Informatica Economica, A.S.E. Bucuresti,

Lect. Ileana TANASE

Universitatea "Constantin Brâncoveanu" 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). Entitatile si interogările sunt reprezentate de vectori, fiecare element al acestora fiind dat de un numar precis ce reprezinta valoarea unui termen index asociat la entitate.

Multimea completa a valorilor dintr-un vector descrie pozitia entitatii sau interogării în spatiu. Între entitati si interogari se pot stabili similitudini bazate pe compararea vectorilor atasati. Ca o masura de similitudine se poate considera atât produsul intern a doi vectori, cât si cosinusul dintre acestia. Aceasta interpretare geometrica a cautarii este destul de intuitiva si reprezinta baza pentru operatiile de cautare a entitatilor. Modelul vectorial se concentreaza pe potrivirea dintre interogari si entitati, iar cautarea entitatilor nu tine seama de relatiile care exista între acestea. Oricum, 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 eficienta de cautare mai mare. Teoria clusterilor se bazeaza pe relatia de similitudine dintre entitati si a fost elaborata mult mai tarziu de Willett în 1988.

Configurarea spatiului de entitati

Regasirea elementelor reprezinta caracteristica principala a organizarii datelor si, în consecinta, gasirea unui sistem optim de indexare este foarte necesara. Matematic se încearca obtinerea unui vocabular de indexare cât mai bun, prin calculul densitatii spatiului S.

Consideram entitatile E_i , fiecare identificata prin unul sau mai multi termeni 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 spatiului este identificat prin trei 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 .

$$E_i = (e_{i1}, e_{i2}, \dots, e_{it}), \quad (1)$$

unde e_j reprezinta valoarea termenilor index T_j .

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

$$Q = (q_a, q_b, \dots, q_r) \quad (2)$$

sau

$$Q = (q_a \text{ and } q_b) \text{ or } (q_c \text{ and } q_d \text{ and } \dots) \text{ or } \dots \quad (3)$$

unde q_k reprezinta termenii asociati interogării Q .

O reprezentare mai formală a vectorilor din ecuațiile (1), (2) și (3) este obținută încluzând în fiecare vector pentru toate componentele o valoare care să creeze distincții între termeni. Dacă w_{ek} și w_{qk} reprezintă valoarea termenului t_k în entitatea E sau în interogarea Q , atunci vectorii asociați pot fi scrși sub formă:

$$\begin{aligned} E &= (e_1 \cdot w_{e1}; e_2 \cdot w_{e2}; \dots; e_t \cdot w_{et}) \quad \text{și} \\ Q &= (q_1 \cdot w_{q1}; q_2 \cdot w_{q2}; \dots; q_t \cdot w_{qt}) \end{aligned} \quad (4)$$

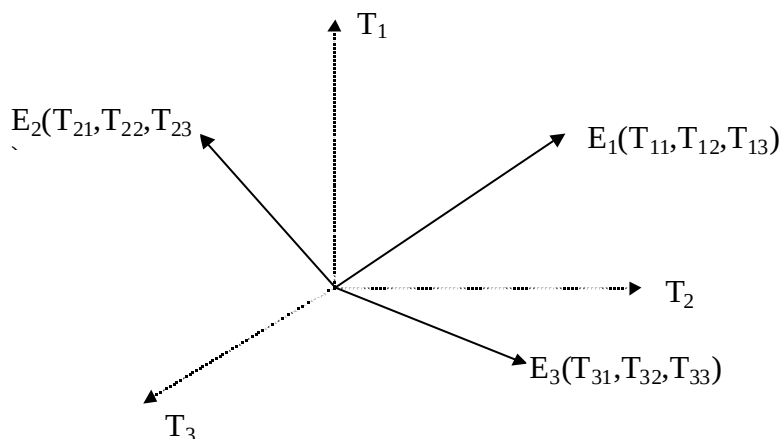


Fig. 1 - Reprezentarea vectorilor în spațiul entitatilor

Vom presupune că w_{ek} (sau w_{qk}) sunt egale cu 0, dacă termenul k nu este asigat la entitatea E (sau interogarea Q) și w_{ek} (sau w_{qk}) sunt egale cu 1 pentru termenii asigati.

Coeficientul de similitudine

Pentru două entități cu vectorii index asociați 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 măsură de similitudine $s(E_i, E_j)$ ar putea fi produsul intern al celor doi vectori

$$\langle E_i, E_j \rangle = \sum_{k=1}^t w_{ik} \cdot w_{jk}$$

sau alternativ unghiul dintre doi vectori

$$\cos(E_i, E_j) = \frac{\langle E_i, E_j \rangle}{\|E_i\| \cdot \|E_j\|}$$

Dacă termenii asociați celor doi vectori sunt identici, unghiul va fi zero, iar similitudinea maximă.

Analog se poate calcula și o valoare de similitudine interogare-entitate obținută tot ca produsul intern al celor doi vectori, sau ca o valoare de cosinus a vectorilor.

$$s(Q, E) = \sum_{k=1}^t w_{qk} \cdot w_{ek}$$

Dacă valorile termenilor index sunt restricționate la 0 și 1, produsul intern a doi vectori măsura numărul de termeni care sunt împreună asigati la ambii vectori în același timp.

În practică, s-a dovedit util să se asigure un grad mai mare de discriminare între termenii asigati pentru reprezentarea conținutului. De aceea, sunt permise nu numai valori de 0 și 1 pentru valorile termenilor, ci și valori care variază între 0 și 1. Asignarea unor valori apropiate de 1 este folosită pentru termenii cei mai importanți, iar valori apropiate de 0 sunt folosite pentru termenii cu o mai mică importanță. În aceste circumstanțe ar putea fi utilă folosirea vectorilor normalizați. Factorul

de norma-lizare ar putea fi $\frac{w_{ek}}{\sqrt{\sum_{\text{vector}} (w_{ei})^2}}$

pentru enti-tati , si $\frac{w_{qk}}{\sqrt{\sum_{\text{vector}} (w_{qi})^2}}$ 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:

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

entitati si, analog, 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 uni-tate si înfasuratoarea spatiului. Doua enti+-tati 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 vec-torii corespondenti. Similitudine mare im-plica distanta mica între puncte si invers.

Problema ar putea fi privita si invers: decât sa reprezentam optim entitatile într-un spa-tiu dat, mai bine am corecta un spatiu astfel încât sa devina optim pentru entitatile date.

Daca nu se cunoaste nimic special despre elementele colectiei s-ar putea presupune ca spatiul optim de entitati este unul unde entitatile sunt grupate daca furnizeaza ace-lasi raspuns la o interogare, asigurând ca vor fi încarcate împreuna. În mod contrar, elementele singulare care nu au simili-tudini cu alte elemente sa apara bine sepa-rate în spatiul de entitati. O astfel de situa-tie este aratata în figura 2, unde distanta dintre doua reprezentari x a doua entitati este invers legata de similaritatea dintre vectorii index corespunzatori.

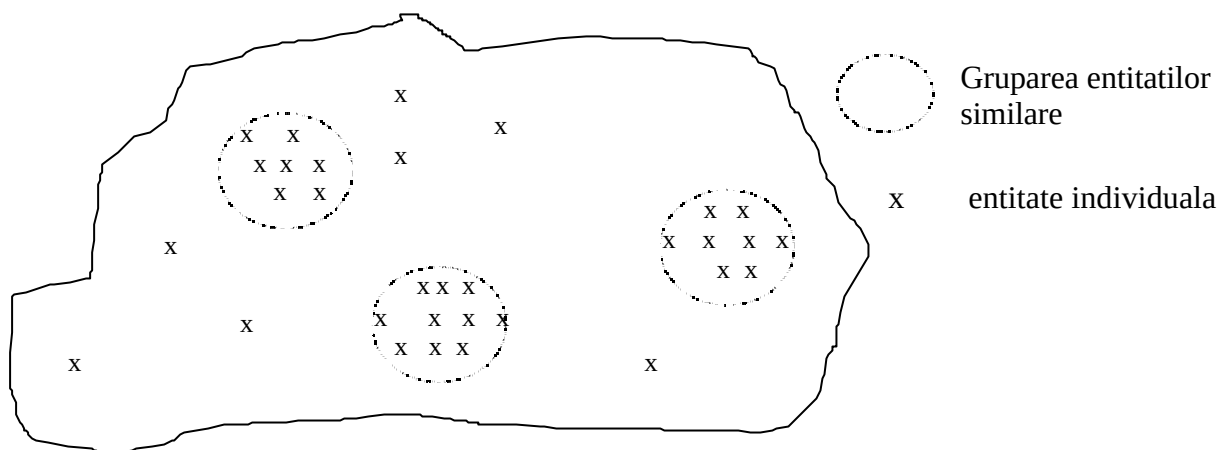


Fig. 2 - Spatiul ideal de entitati

Deși configurarea entitatilor poate fi într-adevar cea optimă, în practică nu există o cale de realizare actuală a unui astfel de spațiu, deoarece în timpul procesului de indexare este dificil de anticipat care va furniza în timp

aproape același lucru.

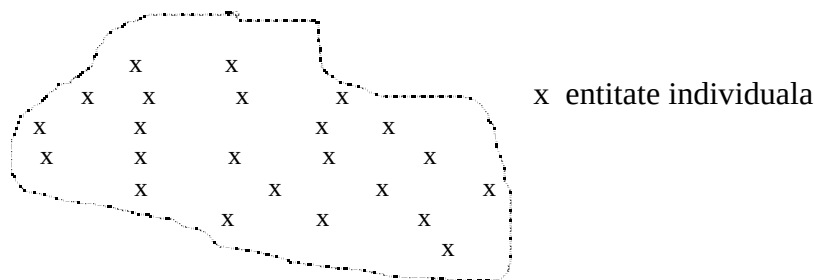


Fig. 3 - Spatiul cu separare maxima între perechi de entitati

Deoarece acest spațiu optim este foarte greu de obținut, s-a încercat o separare a tuturor entitatilor existente în spațiu (figura 3). Acest lucru este obținut dacă similitudinea dintre entități este mică. În acest sens s-a considerat funcția

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

care se dorește a fi minimizată.

$s(E_i, E_j)$ este similitudinea dintre entitățile E_i și E_j .

Se poate observa că atunci când funcția din exemplul (1) este minimizată, media similitudinilor dintre perechile de entități este cea mai mică, astfel garantând că fiecare entitate

poate fi obținută când s-a localizat suficient de aproape de o interogare utilizator "fără a fi nevoie să îi încărcăm și vecinii". Aceasta asigură o înaltă precizie de căutare și da posibilitatea încărcării unei entități relevante fără să încărcăm un număr de entități nerelevante din vecinătatea sa. Aceasta reprezentare este foarte bună atunci când interogarea are ca rezultat o entitate, sau entități aflate în aceeași arie. În cazul în care interogarea este formată din mai multe entități relevante dispersate în spațiu, încărcarea lor implică încărcarea unui număr mare de entități nerelevante. Pe de altă parte, nici funcția din exemplul (1) nu este prea ușor de calculat,

ținând cont ca sunt necesare n^2 comparari, pentru o colecție de n entități.

Din aceasta rațiune, un spațiu de entități care să grupeze entitățile în clase este demn de luat

în considerare. Fiecare clasă existentă este prezentă și cu un centru propriu.

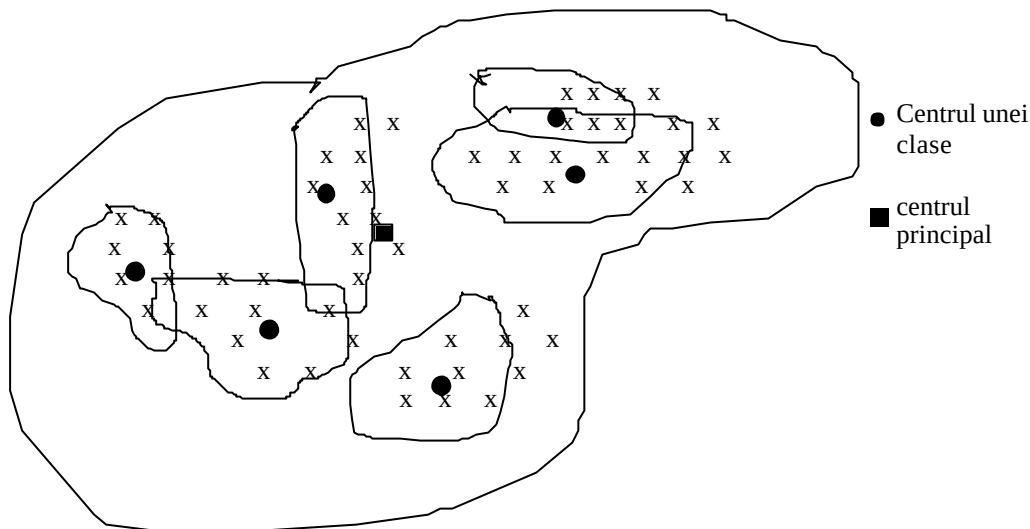


Fig. 4 - Clasele cu centre

Un astfel de spațiu de entități separate în clase arătat în figura 4, unde entitățile care aparțin unei clase sunt grupate, iar centrul fiecărei clase a fost reprezentat prin puncte negre.

Pentru o clasă de entități K care comprimă m entități, centrul clasei K poate fi definit ca fiind media aritmetică a componentelor tuturor entităților (reprezentate vectorial).

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

unde c_j este o reprezentare vectorială pe componente, iar m este numărul entităților din clasa j .

Corespunzător centrului fiecărei clase se poate defini și un centru principal, ca fiind "centrul centrelor". Acest centru principal poate fi obținut din centrele claselor în aceeași manieră în care centrele claselor sunt calculate pentru entitățile individuale. Deci centrul principal va fi pur și simplu media aritmetică a centrelor claselor. Notația folosită pentru centrul principal este c^* .

Într-un spațiu entitate de clase, măsura densității spațiului compusă din suma tuturor similarităților entităților pereche (ecuația 1), poate fi înlocuită ca suma tuturor coeficienților de similitudine dintre fiecare entitate și centrul principal.

$$Q = \sum_{i=1}^n s(c^*, E_i)$$

Dacă pentru funcția din exemplul (1) erau necesare n^2 comparări, evaluarea funcției Q se poate face doar cu n comparări.

Concluzii

Gruparea entităților în clase trebuie făcută după criterii precise. Obținerea unui tip de grupare așa cum se arată în spațiul separat din figura 3 contribuie la o bună performanță de regasire. Dacă ținem seama că entitățile aflate în apropierea unui aceluși centru de clasă arată caracteristici de relevanță identice (respectând majoritatea interogărilor utilizator)

atunci cea mai buna performanta de regasire ar trebui sa fie obtinuta intr-un spatiu de clase individuale compacte, 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 micșorata.

Observatie: clasele sînt separate bine unele de altele iar intr-o clasa entitatile sunt apropiate.

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 different indexing methods on the document space configuration, Cambridge, England, 1987
4. van Rijsbergen C.J., Information retrieval, Butterworths, 1988