Modelul vectorial de organizare a colectiilor de date (II)

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

Modelul vectorial este conceput pe baza unui spatiu n-dimensional în care interogarile si entitatile sunt reprezentate prin vectori ale caror elemente sunt corespunzatoare termenilor index asociati. Valoarea unui sistem de indexare este exprimata în functie de densitatea acestui spatiu care va fi utilizata pentru gasirea unui vocabular optim de indexare pentru o colectie de entitati. Rezultatele evaluarii vor fi aratate demonstrând necesitatea alegerii acestui model.
Cuvinte cheie: *entitate spatiu model vectorial, index.*

Corelarea performantei între densitatea spatiului si indexare

Ideea initiala a modelului vectorial a fost de a reprezenta optim entitatile într-un spatiu dat. Relatia dintre configurarea spatiului si performanta de regasire poate fi considerate si dintr-un punct de vedere opus. În loc sa alegem analiza de entitati si sa indexam testând efectul pe densitatea spatiului entitate este posibil sa schimbam artificial configurarea spatiului entitate pentru a stabili daca s-au produs schimbarile asteptate în regasirea datelor.

Criteriul de densitate de spatiu arata ca o colectie de entitati mici si strâns grupate, cu mari separari între clase, produc cea mai buna performanta. Reversul este adevarat pentru clase mari neomogene, bine separate.

Pentru a realiza o îmbunatatire în performanta ar parea suficient:

a) Sa marim similitudinea dintre vectorii entitatilor grupate în aceeasi clasa. Acest efect este realizat prin accentuarea termenilor care sunt unici numai anumitor clase sau termeni care au frecventa de întâlnire pe cluster puternic nesimetrica (ceea ce înseamna ca se întâlnesc cu frecvente mari în unele clase si cu frecvente mult mai mici în altele).

b) Sa descrestem similitudinea dintre diferitele clase sau centrele acestora. Acest efect se produce prin neaccentuarea termenilor care se întâlnesc în multe clase diferite.

Pentru a îndeplini transformarile necesare este nevoie sa introducem doi parametri:

NC(k)=numarul claselor în care termenul k se întâlneste (un termen se întâlneste într-o clasa daca este asignat la cel putin o entitate din aceea clasa);

CF(k,j)=frecventa claselor cu termenul K inclus în clasa j, adica numarul entitatilor din clasa j în care termenul k se întâlneste.

Pentru o colectie aranjata în *p* clase, frecventa medie <CF(k)> poate fi definita astfel:

$$< CF(k) >= \frac{1}{p} \sum_{j=1}^{m} CF(k, j)$$

Consideram:

$$F_1 = | <CF(k)-CF(k,j)> |$$

Pe de alta parte, un factor F_2 invers lui NC(k), $F_2=1/NC(k)$ poate fi utilizat pentru a reflecta raritatea termenului k. Daca se multiplica ponderea fiecarui termen k în fiecare clasa j printr-un factor proportional cu $F_1 \bullet F_2$ atunci spatiul document se raspândeste. Daca se multiplica cu un factor proportional cu $1/ (F_1 \bullet F_2)$ atunci spatiul va fi comprimat.

Experimentele efectuate în acest sens au aratat ca modificarea valorii termenilor printr-un factor $F_1 \bullet F_2$ produce efectul anticipat: similitudinea dintre documentele aflate în aceeasi clasa este mare, în timp ce similitudinea dintre centrele claselor s-a micsorat. Daca X masoara similitudinea dintre entitatile din aceeasi clasa, iar Y similitudinea dintre centrele claselor, densitatea spatiului poate fi calculata ca fiind

Y/X. Densitatea spatiului este mai mica si media performantelor de regasire creste cu câteva procente.
Rezultate corespunzatoare au fost obtinute si prin transformarea cu un factor 1/($F_1 \bullet F_2$). În acest caz, similitudinea dintre entitatile din aceeasi clasa descreste, în timp ce similitudinea dintre centrele claselor creste. Densitatea spatiului (Y/X) se va mari si acest spatiu al entitatilor produce pierderi în recuperarea si precizia de cautare.
Luate împreuna, aceste rezultate indica:
1. performanta de regasire si densitatea spatiului par sa fie într-o relatie inversa, în sensul ca indexarea eficienta se asociaza cu spatii separate;
2. modificarile generate artificial în densitatile spatiilor par sa produca schimbarile anticipate în performanta.

Modelul valorilor diferentiate
Acest model se bazeaza pe valoarea termenilor index prin adaugarea unui nou termen index la colectia existenta si prin studierea diferentei care se produce între similaritati înainte si dupa adaugarea noului termen.
Un termen index "bun" este cel care are o valoare de diferentiere superioara, capabil sa descreasca similitudinea dintre entitati când este asociat la o colectie, asa cum se arata in figura 1. Analog se obtine notiunea de termen "nepotrivit" care produce o diferentiere slaba.

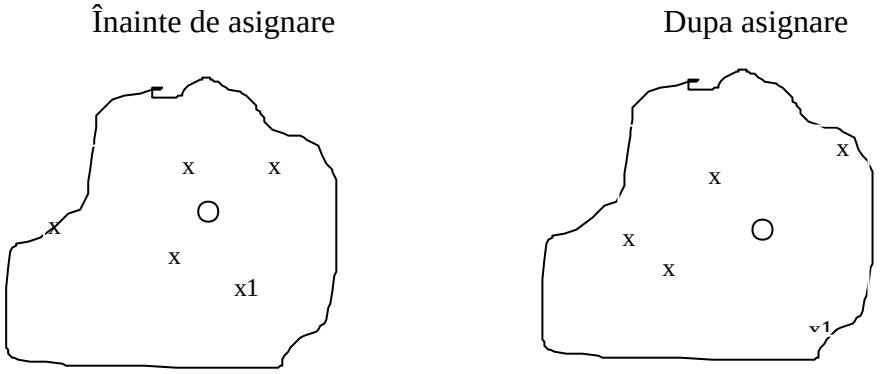

Fig.1. Adaugarea unui termen "bun"

Pentru a masura capacitatea de diferentiere a unui termen este suficienta luarea în considerare a densitatii spatiului înainte si dupa adaugarea unui termen particular. Densitatea spatiului poate fi calculata cu ajutorul functiei Q, definita astfel

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

unde n = numarul de entitati din cadrul colectiei, E_i=entitate, c^*=centrul principal al claselor
Contributia unui termen k dat la densitatea spatiului poate fi stabilita prin calculul functiei:

$$DV_k = Q_k - Q \quad (1)$$

unde Q_k este cel mai compact spatiu-entitate cu termenul k sters din toti vectorii entitate. Daca termenul k este bun discriminator, Q_k-Q>0, altfel se obtine o slaba diferentiere când Q_k-Q<0.

Concluzii:
Pentru a realiza o mai buna performanta este necesar:
a) Sa marim similitudinea dintre vectorii entitatilor grupate în aceeasi clasa. Acest efect este realizat prin accentuarea termenilor care sunt unici numai anumitor clase sau a termenilor care au frecvente mari în anumite clase si frecvente mici în altele.
b) Sa descrestem similitudinea între clase sau centrele acestor clase. Acest efect se produce prin neaccentuarea termenilor care se întalnesc în multe clase diferite.

Generarea interogarilor optime
Problema poate fi abordata si din punctul de vedere al interogarilor. Anumite prelu-

crari asupra acestora (adaugare sau stergere de termeni) contribuie la o mai mare claritate si la o restrângere a cautarii entitatilor care sunt furnizate drept raspuns. Se porneste de la premisa ca similitudinea interogare-document este data de produsul intern al vectorilor corespunzatori.

În acest sens, o importanta deosebita au studiile facute de Rocchio în 1978 cu privire la gasirea unei formule pentru a asigura optimizarea interogarilor. Acest studiu se bazeaza pe vectorii entitate si pe identificarea termenilor relevanti si nerelevanti din cadrul acestor vectori.

$Q_{opt}=1/n$

$$Qopt=\frac{1}{n}\sum_{elem.relevante}\frac{Ei}{|Ei|}-\frac{1}{N-n}\sum_{elem.nerelevante}\frac{Ei}{|Ei|} \quad (2)$$

unde E_i= vectorii document, $|E_i|$=norma euclidiana a vectorului si

$$|Ei|=\sqrt{\sum_{k=1}^{t}(wci)^2}$$

N=dimensiunea colectiei, n=numarul de documente relevante din cadrul colectiei.

Aceasta formula nu poate fi utilizata în practica ca formulare a unei interogari initiale, deoarece nu se poate cunoaste de la început care sunt toate documentele relevante si care nu. Oricum ecuatia (2) este de un real folos, pe baza ei ajungându-se la o iteratie pentru considerarea unei interogari optime. Suma tuturor documentelor relevante si nerelevante normalizate din expresia (2) este înlocuita de o suma cunoscuta inital pentru documentele relevante (n_1) si nerelevante (n_2).

$$Q_1=Q_0+\frac{1}{n_1}\sum_{elem.relevantecunoscute}\frac{Ei}{|Ei|}-\frac{1}{n_2}\sum_{elem.nerelevantecunoscute}\frac{Ei}{|Ei|}$$

unde Q_0 = interogarea initiala de la care se porneste optimizarea

Q_1=prima interogare (iteratie)

Aceasta formula poate fi generalizata prin utilizarea unor factori de multiplicare α,β,γ.

$$Q_{i+1}=a\,Q_i+b\sum_{elem.relvante}\frac{Ei}{|Ei|}-g\sum_{elem.nerelevante}\frac{Ei}{|Ei|}$$

Metoda de modificare a vectorului interogare descrisa pâna acum este destul de simpla, deoarece valoarea termenului modificat este direct obtinuta din valorile termenilor corespunzatori din documentele cunoscute ca fiind relevante sau nerelevante pentru interogarile respective. Procesul de modificare a vectorului standard furnizeaza o metoda buna pentru construirea unei interogari considerate optime.

Desi au existat multe încercari nu s-a reusit pâna acum sa se dezvolte o implementare operationala care sa obtina rezulatele asteptate. Modelul vectorial cauta sa depaseasca limitarile modelului boolean, dar are propriile sale limitari. Cea mai evidenta limitare este aceea ca termenii de indexare folositi pentru a defini dimensiunile spatiului de cautare trebuie sa fie ortogonali, o presupunere care nu se adevereste întotdeauna.

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