Modelarea matematică –fundament al regăsirii informațiilor

Mihaela VOINICU
director adjunct Biblioteca Județeană ”Dinicu Golescu” Argeș

Regăsirea informațiilor (Information Retrieval) este un subiect strâns legat de Știința Informării și de Biblioteconomie. În cadrul acestui subiect, modelarea matematică a sistemelor de regăsire a informațiilor reprezintă un domeniu aparte, despre care se vorbește și se studiază mai degrabă în facultățile de informatică și calculatoare, în cadrul cursurilor de inteligență artificială.

În încercarea de a defini și aborda Știința Informării precum orice altă știință, fie ea exactă sau nu, construirea unei baze teoretice fundamentale ar trebui să fie o preocupare permanentă a specialiștilor. Un simplu exemplu al aplicabilității dezvoltărilor teoretice din domeniul bibliometriei este edificator: algoritmul de regăsire a informației din spațiul web, PageRank, pe care s-a bazat ideea dezvoltată de Lawrence Page și Sergey Brin, în anii 1995-1998, în cadrul unui proiect de cercetare la Universitatea Stanford, a condus la apariția motorului de căutare *Google*. Esența ideii inițiale a acestui algoritm se baza pe luarea în considerare a importanței paginii web de unde se face referirea, idee preluată și adaptată din metoda analizei citărilor.

În procesul de regăsire a informațiilor distingem, din perspectiva sistemelor care interacționează, două componente: cea a intervenției umane directe (identificarea unei nevoi de informare, formularea unei cereri de informare și a unei strategii de căutare) și cea a intervenției automatizate (rezolvarea ecuației de căutare și determinarea relevanței răspunsului).

În contextul existenței unor baze de date eterogene și foarte mari în care se caută răspunsul la o interogare specifică, ne dăm seama că toate componentele procesului automatizat de regăsire și de returnare a datelor relevante în scopul obținerii unor informații (informațiile fiind privite ca date puse într-un anumit context) trebuie tratate cu maximă atenție și rigurozitate.

Continuând logica argumentării putem spune că în procesul automatizat de rezolvare a ecuației de căutare și determinare a relevanței răspunsului, esențial este modelul matematic care stă la baza sistemului.

Un model matematic este o încercare de transpunere și simplificare a realității cu ajutorul unor expresii matematice. El poate descrie mai bine o anumită entitate decât o face o descriere verbală iar construcția unui astfel de model este o condiție *sine qua non* pentru implementarea unui sistem automatizat de gestiune a documentelor.

Avantajele oferite de modelarea matematică sunt evidente, în contextul actualelor tehnologii informatice, permițând exploatarea teoriilor matematice cunoscute și aplicarea lor uniformă și sistematică asupra unui volum mare de date, precum și studierea efectelor pe care modificarea unor parametrii/date de intrare le au asupra rezultatelor/ datelor de ieșire.

Modelele matematice destinate regăsirii informațiilor au cunoscut evoluții. Fixarea în timp a întregului proces de dezvoltare este utilă și necesară în cunoașterea imaginii de ansamblu și în analiza legăturilor, interdisciplinare, dintre diverse domenii ale cunoașterii. Încercarea de reconstituire cronologică pe care o prezentăm este deschisă corecțiilor și modificărilor, având în vedere faptul că, uneori, aspecte sau părți ale modelelor dezvoltate au fost publicate sau prezentate parțial de către autori, anterior datei menționate mai jos. Principalele repere cronologice ale acestei evoluții sunt:

-Modelul Boolean (Lancaster & Fayen,1950) model bazat pe algebra booleană care operează cu operatorii .AND. și .OR. aplicați asupra termenilor asociați documentelor și interogării pentru a returna o listă de răspuns.

-Modelul Vectorial (Salton &Wong & Yang, 1970) asociază documentele și interogarea cu un vector, calculează similaritatea documentelor pe baza unei măsuri de similaritate.

-Modelul Probabilistic (Jones &al., 1976) bazat pe calcularea probabilității posterioare - probabilitatea ca un nou document să aparțină unei anumite colecții, cunoscând probabilitatea anterioară, probabilitate calculată pe un set de date de antrenament.

-Modelul Boolean Extins (Salton, 1983) combină modelul boolean clasic cu cel vectorial.

-Modelul bazat pe logica fuzzy (1984) introduce noțiunea de similaritate graduală între document și interogare.

-Modelul Indexării Semantice Latente (Deerwester & al., 1990) identifică relații între termeni și concepte pornind de la ipoteza că termenii utilizați în același context au sensuri/înțelesuri similare, stabilind șabloane în vederea extragerii conceptelor dintr-un text.

-Modelul "Mașini cu Suport Vectorial" (Vapnik, 1995) apelează la reprezentarea vectorială a documentelor și construiește un hiperplan sau o mulțime de hiperplane într-un spațiu multidimensional astfel încât să separe optim membri ai claselor diferite.

-Modelul lingvistic (Ponte & Croft, 1998) are la bază ideea de a extrage descrierea lingvistică dintr-un text și a estima probabilitatea ca acel document să corespundă unei interogări. Se are în vedere componenta semantică a unui text care este mult mai cuprinzătoare decât o simplă analiză sintactică.

În cadrul acestei enumerări mai amintim și încercările de modelare bazate pe teoria graph-urilor, care în fapt își au sursa în dezvoltarea teoriei hipertextului. În esență această teorie poziționează documentele și interogarea în vârfurile (nodurile) unui graph iar muchiile graph-ului constituie legăturile care conectează aceste documente. Structura astfel rezultată poate fi ierarhică sau liniară, legăturile pot fi unidirecționale sau bidirecționale. Rețeaua rezultată poate fi comparată cu modalitatea în care creierul uman poate realiza asociații între idei și concepte.

Dintre aceste modele dezvoltate s-a detașat cu claritate, prin prisma aplicabilității în biblioteconomie, modelul boolean.

Prezentăm în continuare, într-o manieră ce nu apelează la matematica formalizată, acest model care a dovedit, de-a lungul timpului, că nu și-a epuizat toate resursele deoarece a avut parte de noi dezvoltări și abordări teoretice dintre care amintim: modelul bazat pe logica fuzzy și modelul boolean extins

Modelul boolean se bazează pe algebra booleană și identifică trei tipuri de relații de dependență cu ajutorul operatorilor AND, OR și NOT. Prin aplicarea unei ecuații de interogare cu ajutorul acestor operatori asupra descriptorilor atașați unui document, modelul evaluează ce descriptori sunt prezenți (1) sau absenți (0) dintr-un document.

Se observă că în cazul folosirii operatorului .AND. se va returna o listă care va avea lungimea egală cu minimul ocurențelor termenilor în text.

În cazul folosirii operatorului .OR. se va returna o listă care va avea lungimea egală cu maximul ocurențelor termenilor în text.

Operatorul .NOT. este deosebit de puternic. Folosirea lui presupune constrângeri suplimentare și trebuie folosit cu discernământ deoarece poate reduce nejustificat lista de răspuns și implicit nereturnarea unor documente relevante.

Modelul bazat pe logica fuzzy (teoria vagului). Logica fuzzy permite o interpretare mai flexibilă a noțiunii de apartenență. Dacă în algebra booleană un obiect (document) putea să aparțină (1) sau nu (0) unei mulțimi, în logica fuzzy se definește un continuum, în intervalul [0,1], astfel încât un obiect (document) poate să aparțină unei mulțimi în grade diferite în acest interval (în sensul în care în intervalul [0,1] valoarea 0,01 este mai "apropiată" de 0 decât valoarea 0,1). În acest fel un obiect (document) nu este automat exclus din lista de răspuns la o interogare. El este plasat în apropierea/ vecinătatea listei de răspuns, fapt similar cu folosirea operatorului .NEAR. sau .NEXT. Astfel, în urma unei interogări, se pot obține, pe baza

operațiunilor cu mulțimi fuzzy (reuniune, intersecție, inferență), liste de documente ordonate după relevanță.

În practică, decizia stabilirii importanței unui descriptor atașat unui document se dovedește a fi o practică neuniformă și subiectivă.

Modelul boolean extins este o combinație între modelul boolean clasic și modelul vectorial prin introducerea unei ecuații de ponderare a termenilor și de calcul a similarității în procesul de stabilire a relevanței și returnare a listei de răspuns la o interogare.

În VSM (Vector Space Model) fiecare document este reprezentat ca un vector de caracteristici, a cărui lungime este egală cu numărul de descriptori ai documentului din colecție. Fiecare descriptor are asociată o pondere. În cel mai simplu caz aceste ponderi pot fi binare, indicând prezența sau absența termenului în document.

O altă metodă de ponderare a termenilor este folosirea frecvenței de întâlnire a termenului în document, bazată pe statistica aparițiilor termenului în document. O descriere a acestei metode a fost făcută de Salton și Buckley[1]. Astfel, dacă se notează cu Tf frecvența de întâlnire a termenului în document, iar cu Idf inversul frecvenței de întâlnire a termenului în întreaga colecție, atunci:

$$Idf=\log (N/n_k)$$

unde n_k este numărul de documente care conțin termenul, iar N este numărul total de documente din colecție.

Produsul tf*Idf este cea mai cunoscută măsură de ponderare a termenilor. Valoarea acestui produs este mare atunci când un anumit termen se întâlnește de foarte multe ori într-un număr restrâns de documente comparativ cu întreaga colecție.

Revenind la problema reprezentării documentelor sub forma unor vectori apare problema lungimii acestora. Astfel, un document caracterizat prin mai mulți descriptori va avea o lungime mai mare, în timp ce o interogare are, de obicei, o lungime mai scurtă.

Pentru a calcula asemănarea dintre două documente sau dintre un document și o interogare se alege o măsură de similaritate.

Exemple de astfel de măsuri sunt: distanța euclidiană, măsura Jaccard, coeficientul Dice.

Cea mai cunoscută măsură de similaritate este cea cosinus, care măsoară cosinusul dintre doi vectori de caracteristici. Similaritatea cosinus este o metrică normalizată deoarece poate avea valori doar în intervalul [0,1] eliminându-se astfel dependența de lungimea vectorului și implicit de numărul de termeni asociați documentului.

În modelul boolean extins se introduce noțiunea de suprapunere (similaritate) parțială între documente și interogare. Se aplică metodele de ponderare a termenilor din modelul vectorial. Interogarea este formulată cu ajutorul operatorilor booleeni, model considerat mai simplu și mai intuitiv pentru utilizator. În a treia etapă se formulează răspunsul la interogare pe baza similarității între termenii interogării și cei ai documentului, care sunt termeni ponderați. În acest fel un document poate fi parțial relevant dacă există un grad de similaritate între termeni.

Concluzii

Modelul boolean clasic și-a dovedit eficacitatea raportat la nevoile pe care un software integrat de bibliotecă le are. În esență el operează cu seturi de date care sunt ușor de capturat, organizat și procesat. În practică acest model s-a dovedit a fi unul foarte bun, el scanează rapid baze mari de date și procesează rapid o interogare.

[1] http://www.cs.odu.edu/~jbollen/IR04/readings/article1-29-03.pdf

Între dezavantajele pe care le aduce amintim faptul că deși returnează rapid un răspuns la o interogare el nu are capacitatea de a returna o listă ordonată după un criteriu de relevanță (descriptorii atașați au aceeași pondere) și nu poate realiza o corelare la nivel semantic și conceptual între mai mulți descriptori.

Existența unui anumit model matematic la baza unui software de bibliotecă determină tipul de management care se va face în cadrul sistemului: management al datelor și/sau management al informațiilor.

Pentru a trece la un nivel superior de regăsire a informațiilor este necesar a fi transformate datele în informații, fapt ce se poate petrece doar prin contextualizare fațetată, operare cu concepte și ontologii.

Bibliografie

Buraga, Sabin Corneliu, *Semantic web – fundamente și aplicații*, București, Editura MatrixRom, 2004

Le Coadic, Yves-F., *Știința informării*, București, Editura Sigma, 2004

Gorunescu, Florin, *Data Mining. Modele și Tehnici*, Cluj-Napoca, Editura Albastră, 2006

Morariu, Daniel I., *Text mining methods based on Support Vector machine*, Bucharest, Ed. MatrixRom, 2008

Negoiță, Constantin Virgil, *Sisteme de înmagazinare și regăsire a informațiilor*, București, Editura Academiei Republicii Socialiste România, 1970

https://en.wikipedia.org/wiki/Latent_semantic_analysis [accesat septembrie 2015].

Benno Stein, Tim Gollub, Maik Anderka, *Retrieval Models*. [accesat septembrie 2015]. Disponibil pe internet la adresa http://www.uni-weimar.de/medien/webis/publications/papers/stein_2014n.pdf

Truong, Quoc-Dinh; Dkaki, Taoufiq; Mothe, Josiane; Charrel, Pierre-Jean, *Information retrieval model based on graph comparison*. [accesat septembrie 2015]. Disponibil pe internet la adresa http://lexicometrica.univ-paris3.fr/jadt/jadt2008/pdf/truong-dkaki-mothe-charrel.pdf