Optimizarea regăsirii informaţiilor prin modelarea matematică

MIHAELA VOINICU
Biblioteca Judeţeană „Dinicu Golescu" Argeş

Abstract:
The Optimization of Retrieving Information Through Mathematical Modeling

Information Retrieval Systems are software that save and carry out the management of the information contained in documents. Any Information Retrieval System is based on a mathematical model. The main mathematical models on which the Information Retrieval Systems are based on are: the Boolean model, the Vector Space model, the probabilistic model and the linguistic model. In addition to these classical models, in the recent years there have been added the models based on clustering methods, Latent Semantic Indexing and Support Vector Machines. The present article describes the principles underlying these models, the advantages and disadvantages for each model, with a view to clarify the role these models play or may play in the information retrieval process..

Keywords: *information retrieval, Boolean model, Vector Space model, probabilistic model, clustering, latent semantic indexing, support vector machines.*

1. Introducere

Pornind de la realitatea faptului că una dintre fundamentalele funcţii ale bibliotecii este aceea de regăsire a informaţiilor, modernizarea serviciilor de bibliotecă nu poate fi concepută fără un studiu temeinic al principalelor modele matematice ale IR (Information Retrieval) în vederea implementării celor mai viabile soluţii în cadrul softurilor de bibliotecă.

Regăsirea datelor în contextul unui Sistem de Regăsire a Informaţiilor constă, în principiu, în determinarea documentelor unei colecţii care conţin cuvintele cheie din interogarea utilizatorului, dar care nu sunt suficiente, de cele mai multe ori, în furnizarea informaţiei necesare utilizatorului. În general, utilizatorul unui sistem IR este preocupat mai mult de regăsirea informaţiei referitoare la un subiect decât

BDD-A19436 © 2010 Biblioteca Județeană Mureș
Provided by Diacronia.ro for IP 216.73.217.42 (2026-11-15 20:59:08 UTC)

de regăsirea datelor care satisfac o interogare dată. Dată fiind interogarea, scopul cheie al unui sistem IR este regăsirea informaţiei care ar putea fi utilă sau relevantă pentru utilizator. Se pune accentul pe regăsirea **informaţiei**, şi nu pe regăsirea **datelor**.

Regăsirea informaţiilor se bazează în primul rând pe măsurători care se pot face în cadrul sistemului, pe baza modalităţilor de reprezentare a documentelor şi a modului de evaluare a similarităţii documentelor cu alte documente sau cu termenii unei interogări. Un document poate fi caracterizat şi identificat cu ajutorul unui ansamblu de atribute textuale (cuvinte cheie, descriptori) şi paratextuale (autor, titlu, editură, an etc.). Practic, un document este reprezentat de o listă de termeni care apar în el, termeni ce aparţin unui vocabular (controlat sau nu).

Orice Sistem de Regăsire a Informaţiilor are în spatele lui un model matematic pe care se bazează. Un model al unei entităţi este o reprezentare care reliefează caracteristicile entităţii, putând servi ca un substitut. Modelul este întotdeauna o aproximare, o simplificare a realităţii.

Avantajele oferite de modelarea matematică sunt următoarele:

- indică tipul şi volumul datelor de corelat şi de măsurat;
- permite considerarea entităţii ca un tot;
- permite determinarea efectului modificării unor variabile asupra celorlalte variabile;
- permite utilizarea teoriilor matematice cunoscute;
- permite abordarea logică şi sistematică;
- permite utilizarea tehnologiilor informatice şi de comunicaţii.

Prezentarea într-o formă agregată a modelelor matematice ce stau la baza regăsirii informaţiilor se dovedeşte a fi un demers complicat, datorită varietăţii ştiinţelor sau domeniilor ştiinţelor ce tratează acest subiect (matematica, inteligenţa artificială, lingvistica computaţională, ştiinţa informării etc.).

O scurtă trecere în revistă a principiilor ce stau la baza acestor modele, a avantajelor şi dezavantajelor proprii fiecărui model, /fără virgulă va clarifica rolul pe care aceste modele îl au sau îl pot avea în dezvoltarea Sistemelor de Regăsire a Informaţiilor.

BDD-A19436 © 2010 Biblioteca Județeană Mureș
Provided by Diacronia.ro for IP 216.73.217.42 (2026-11-15 20:59:08 UTC)

2. Modele matematice în informatica documentară

Modelul boolean. Este unul dintre primele şi cele mai simple modele de reperare a informaţiilor. Acest model identifică trei tipuri de relaţii de dependenţă, datorită operatorilor booleni AND, OR şi NOT. Pe baza combinării descriptorilor cu ajutorul operatorilor booleni se urmăreşte suprapunerea exactă între document şi interogare. Modelul boolean evaluează ce descriptori sunt prezenţi sau absenţi din document, obţinând astfel o sumă de documente relevante (descriptori prezenţi) şi excluzând documentele irelevante (descriptorii lipsesc din document).

Avantaje: este un model robust, procesează rapid o interogare, scanează bine colecţii mari de documente. Creşterea colecţiei de documente - cum este cazul bibliotecilor - nu reprezintă o problemă pentru acest model.

Dezavantaje: modelul boolean standard nu poate regăsi documentele pe baza relaţiilor semantice sau conceptuale între descriptori; aceştia sunt trataţi ca fiind independenţi unul de celălalt, egali în importanţă. Prin urmare, rezultatul unei interogări nu poate fi ordonat după relevanţă, obţinându-se liste de documente foarte mari şi greu de exploatat.

În acest model de bază nu există conceptul de suprapunere parţială între document şi interogare. Acestei ultime probleme i s-au găsit rezolvări teoretice bazate pe teoria vagului (logica fuzzy). Modelul rezultat este cunoscut şi ca **modelul boolean extins**.

Logica tradiţională consideră că un obiect (în cazul nostru, un document) poate aparţine, sau nu, unei mulţimi. În acelaşi timp, atribuirea rolului de descriptor unui cuvânt dintr-un document poate să fie descrisă într-o manieră vagă, cum ar fi *destul de important* sau *foarte semnificativ*. Rezultatele procesului de regăsire pot fi *foarte relevante* sau *parţial relevante*. În acest, caz problema se rezumă la stabilirea transformării unor astfel de expresii în funcţii de apartenenţă, asociate cu logica vagului.

Logica fuzzy permite o interpretare mai flexibilă a noţiunii de apartenenţă. Astfel, mai multe obiecte (documente) pot aparţine unei mulţimi în grade diferite. Având în vedere aplicaţiile logicii fuzzy în regăsirea documentelor relevante, putem spune că pentru orice interogare există documente care pot fi mai relevante decât altele. Această exprimare lingvistică poate fi exprimată matematic folosind noţiunile fundamentale

BDD-A19436 © 2010 Biblioteca Județeană Mureș
Provided by Diacronia.ro for IP 216.73.217.42 (2026-11-15 20:59:08 UTC)

ce caracterizează mulţimile fuzzy. Lista de documente returnate în urma efectuării unei căutări poate fi asimilată unei submulţimi fuzzy.

Dacă fiecărui termen (descriptor) ce îi este ataşat unui document în procesul de indexare şi prelucrare îi este acordată o pondere, prin care se exprimă în ce măsură acel descriptor este asociat cu documentul în sine în procesul de regăsire a documentelor, se pot obţine liste de documente realizate pe baza operaţiunilor cu mulţimi fuzzy (reuniune, intersecţie, inferenţă). În practică, decizia stabilirii importanţei unui descriptor ataşat unui document se dovedeşte a fi total subiectivă şi neuniformă, fiind raportată la personalitatea bibliotecarului indexator.

Modelul vectorial (Vector Space Model). În acest model, fiecare document este reprezentat ca un vector de caracteristici, a cărui lungime este egală cu numărul de descriptori ai documentului din colecţie. De obicei, aceşti descriptori sunt termeni care sunt extraşi din document.

După faza de extragere a termenilor care caracterizează un document, urmează etapa de *ponderare a termenilor*; acestor termeni le sunt atribuite ponderi, indicând semnificaţia lor în caracterizarea documentului. Aceste ponderi pot fi binare, indicând existenţa (1) sau nu (0) a termenilor în document.

O altă metodă de ponderare a termenilor, mult mai răspândită, este folosirea frecvenţei de întâlnire a termenului în document sau un algoritm aparţinând familiei Tf*Idf. Frecvenţa de întâlnire a termenului este bazată pe statistica apariţiilor termenului în document şi este cel mai simplu mod de a atribui ponderi unui termen. Tf*Idf este o măsura folosită în colecţiile de documente care favorizează acei termeni care se regăsesc frecvent în documente relevante, dar sunt puţin frecvenţi în colecţie ca întreg. Această situaţie se întâlneşte mai ales în cazul termenilor strict ştiinţifici, de foarte îngustă specializare. Tf reprezintă frecvenţa de întâlnire a termenului în document, iar Idf este inversul frecvenţei de întâlnire a termenului în întreaga colecţie.

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

unde n_k este numărul de documente care conţin termenul, iar N este numărul total de documente.

BDD-A19436 © 2010 Biblioteca Județeană Mureș
Provided by Diacronia.ro for IP 216.73.217.42 (2026-11-15 20:59:08 UTC)

Pentru calculul asemănării dintre două documente urmează alegerea unei măsuri de similaritate, dintre care cele mai cunoscute sunt măsura cosinus, coeficientul Jaccard şi coeficientul Dice.

În literatura de specialitate sunt menţionate şi alte încercări de ponderare a termenilor ce caracterizează un vector document. Astfel, se pot acorda ponderi diferenţiate în funcţie de locul în care se găsesc descriptorii în cadrul documentului (în titlu, în abstract, în conţinut sau la concluzii) sau în funcţie de modul în care sunt scrişi (cu litere aldine sau italice), deoarece se presupune că numai termenii semnificanţi sunt notaţi cu astfel de caracteree.

Între *avantajele* acestui model, cel mai important este faptul că rezultatele unei interogări pot fi ordonate după rang, în funcţie de ponderea importanţei termenilor.

Între *dezavantaje* amintim complexitatea calculului similarităţii care creşte odată cu numărul termenilor din document, respectiv cu lungimea vectorului; modificarea unui termen presupune recalcularea tuturor vectorilor şi a similarităţii dintre documente, respectiv documente şi interogare.

Modelul probabilistic. Acest model se bazează pe calcularea probabilităţii ca un document *d*, ce aparţine colecţiei de documente *C*, să fie relevant pentru utilizator.

La baza acestui model stă Algoritmul Naive Bayes care se foloseşte pentru a clasifica date neetichetate; acest lucru se realizează cu ajutorul unor estimări, folosindu-se date de antrenare etichetate. Conform teoremei Bayes, se poate calcula probabilitatea ca un document să aparţină unei categorii; adică se poate obţine probabilitatea posterioară dacă se cunoaşte probabilitatea anterioară. Estimarea acestor probabilităţi se face prin măsurarea frecvenţei de apariţie a cuvintelor într-un set de documente de antrenare. Această metodă se bazează pe simplificarea supoziţiilor (independenţa reciprocă a atributelor ce caracterizează documentul) de unde vine şi denumirea de Naive. Presupunerea că fiecare cuvânt sau descriptor al unui document este independent (are o apariţie neafectată) de prezenţa sau absenţa oricărui alt cuvânt sau descriptor din document stă la baza formulării matematice a acestei teorii.

BDD-A19436 © 2010 Biblioteca Județeană Mureș
Provided by Diacronia.ro for IP 216.73.217.42 (2026-11-15 20:59:08 UTC)

Fie $d \in D$ unde D este mulţimea documentelor; fie $C=\{c_1,c_2,...c_j\}$ mulţimea claselor (categoriilor sau etichetelor).

Probabilitatea ca un document *d* să aparţină clasei *c* se calculează ca:

$$P(c|d) = P(c) \prod_{i=1}^{n_d} P(t_i \mid c)$$

unde $P(t_i|c)$ este probabilitatea ca termenul t_i să aparţină unui document din clasa *c*. $P(t_i|c)$ este interpretat ca o măsură a evidenţierii a cât de mult contribuie t_i la desemnarea clasei *c*. P(c) este posibilitatea apriorică ca un document să aparţină clasei *c*. $\{t_1, t_2...t_{nd}\}$ sunt atributele (descriptorii) documentului *d*, parte a vocabularului folosit pentru clasificare, iar n_d este numărul atributelor documentului *d*.

Deoarece nu se cunosc valorile reale ale parametrilor pentru P(c) şi $P(t_i|c)$, notăm cu $\widehat{P}(c)$ valoarea estimată din mulţimea de antrenament.

$$\widehat{P}(c) = \frac{N_c}{N}$$

N_c este numărul de documente din clasa *c*, iar N este numărul total de documente.

Astfel, estimăm:

$$\widehat{P}(d \mid c)) = \frac{1 + T_{ct}}{V + \sum_{t'=1}^{V} T_{ct'}}$$

unde T_{ct} reprezintă numărul de apariţii al termenului t în documentele clasei c, iar V este numărul de termeni din vocabular.

Modelul probabilistic de regăsire a documentelor se bazează pe patru componente:[1]

- clasa (variabila dependentă a modelului) - care este o variabilă categorială, reprezentând eticheta sub care va fi cunoscută data clasificată;

[1] C. V. Negoiţă, ***Modelarea matematică în documentare***, Bucureşti, Institutul Central de Documentare Tehnică, 1971.

BDD-A19436 © 2010 Biblioteca Județeană Mureș
Provided by Diacronia.ro for IP 216.73.217.42 (2026-11-15 20:59:08 UTC)

- predictorii (variabilele independente ale modelului) - reprezentaţi de caracteristicele datelor ce urmează a fi clasificate, pe baza cărora se face clasificarea;
- mulţimea de antrenament (învăţare) - care este reprezentată de setul de date care conţine valori pentru cele două componente anterioare şi care este utilizată pentru antrenarea modelului ca să recunoască clasa corespunzătoare, pe baza predictorilor disponibili;
- mulţimea de testare - conţine date noi care vor trebui clasificate de modelul construit şi, astfel, se va putea evalua acurateţea clasificării, adică performanţa modelului.

Modelul probabilistic de regăsire a informaţiilor relevante operează recursiv şi necesită ca algoritmul din spatele metodei să fie iniţializat cu parametri care, apoi, prin calcul iterativ, sunt îmbunătăţiţi, astfel încât să se obţină un scor ce desemnează relevanţa probabilă.

Principalele *avantaje* ale acestei metode care se bazează pe Algoritmul Bayes Naive sunt: este robustă în ceea ce priveşte izolarea zgomotului din date; în cazul valorilor lipsă, ignoră instanţa în timpul estimării probabilităţilor; este robustă la atribute irelevante.

Între *dezavantajele* acestei metode amintim: complexitatea metodei creşte rapid; pe măsura creşterii colecţiei de documente, scanarea acestei colecţii este din ce în ce mai greoaie; presupune unele metode simplificatoare, precum independenţa termenilor.

Luarea în considerare a unui istoric al interogărilor utilizatorului, în cadrul acestui model, poate îmbunătăţi stabilirea parametrilor iniţiali ai mulţimii de antrenament.

Modelul lingvistic. Din punct de vedere ştiinţific, limbajul natural (uman) constituie obiectul de cercetare a numeroase discipline şi, în primul rând, al lingvisticii sau „ştiinţei limbii." Acelaşi obiect de investigaţie interesează însă şi filozofia, psiholingvistica, lingvistica matematică, lingvistica computaţională, prelucrarea limbajului natural.

Analiza semanticii latente (LSA-Latent Semantic Analysis) este atât o o teorie, cât şi o metodă utilizată pentru extragerea şi reprezentarea legăturilor dintre cuvinte şi înţelesul contextual al acestora prin metode computaţionale, aplicate corpusurilor mari de text.

BDD-A19436 © 2010 Biblioteca Județeană Mureș
Provided by Diacronia.ro for IP 216.73.217.42 (2026-11-15 20:59:08 UTC)

Această metodă foloseşte o matrice rară care conţine pe coloane documentele în care se face căutarea, iar pe linii termenii (de obicei termeni trecuţi prin procedura de stemmer - extragerea rădăcinii termenilor) conţinuţi în aceste documente. LSA-ul transformă această matrice a apariţiilor într-o relaţie dintre termeni şi concepte şi o relaţie între acele concepte şi documente. Astfel, termenii şi documentele sunt indirect legaţi prin concepte.

Indexarea semantică latentă se bazează pe aplicarea unei tehnici matematice numite Descompunerea Valorilor Singulare (SVD-Singular Value Decomposition). Aplicarea ei în clasificarea documentelor sau în regăsirea informaţiilor a fost propusă de Deerwester la începutul anilor '90.

În cazul aplicării SVD, matricea iniţială A este descompusă în produsul a trei matrice:[2]

$$A = T_{txn} S_{nxn} (D_{dxn})^T$$

unde t = numărul de termeni, d = numărul documentelor, $n = \min(t,d)$

Matricile T şi D sunt ortogonale. Prin restrângerea matricilor T, S şi D la un rang k, $k < n$, şi recompunerea acestora într-un spaţiu dimensional redus, se obţine matricea:

$$\hat{A} = T_{txk} S_{kxk} (D_{dxk})^T$$

ce reprezintă o aproximare pătratică a lui A de către o matrice de rang k.

Pentru orice document d este înlocuit vectorul document de dimensiune mare cu unul ce exclude termenii eliminaţi prin procesul de SVD. Comparând elementele matricei astfel reconstruită cu matricea iniţială, se observă cum această metodă induce relaţii de similaritate, relaţii ce aproximează judecata umană asupra înţelesului şi similarităţii documentelor.

Între *avantajele* acestei metode se remarcă faptul că LSI poate identifica similaritatea semantică între documente ce aparent nu se aseamănă. Ea dă rezultate bune într-o colecţie cu un vocabular eterogen,

[2] Thomas K Landauer & al, ***Handbook of Latent Semantic Analysis***, New Jersey, Lawrence Erlbaum Associates Inc. Publishers, 2007.

BDD-A19436 © 2010 Biblioteca Judeţeană Mureş
Provided by Diacronia.ro for IP 216.73.217.42 (2026-11-15 20:59:08 UTC)

unde documentele pot folosi termeni diferiţi pentru a face referire la acelaşi subiect.

Între *dezavantaje* amintim: apariţia simultană a termenilor poate induce false rezultate şi o precizie mai scăzută; în colecţii de documente cu un vocabular omogen, rezultatele acestei metode sunt mai puţin folositoare.

Modelul bazat pe clustere. Clusterizarea este una dintre tehnicile fundamentale din domeniul analizei datelor care organizează un set de obiecte dintr-un spaţiu multidimensional, în grupuri coezive, numite clustere. Tehnicile de clustering (grupare spaţială) reprezintă tehnici speciale de aranjare a datelor de intrare pe baza dispunerii spaţiale a vectorilor corespunzători. Pentru a analiza asemănarea/deosebirea dintre elementele unei mulţimi, în vederea grupării, fiecare dintre aceste elemente este definit printr-un vector, ale cărui componente sunt chiar caracteristicile/atributele reprezentative ale vectorului respectiv. În urma procesului de clustering rezultă una sau mai multe clustere (grupe), în funcţie de situaţie, care reprezintă poziţionarea spaţială a proprietăţilor considerate pentru elementele supuse grupării. În interiorul unui asemenea cluster, punctele sunt mai apropiate între ele sau în raport cu un centru comun, decât în raport cu centrele altor grupe.

Un alt gen de clustering este gruparea conceptuală. În cadrul acestui tip de grupare, două sau mai multe elemente aparţin aceleiaşi grupe dacă aceasta defineşte un concept comun tuturor elementelor. Altfel spus, elementele sunt grupate în conformitate cu potrivirea la conceptele descriptive şi nu în conformitate cu măsurile de similaritate.

Procesul de clustering va fi unul de succes dacă, atât similaritatea intra-cluster, cât şi disimilaritatea inter-clustere sunt maxime.

Metodologia clustering are două abordări distincte: clusterizarea ierarhică şi clusterizarea neierarhică/partiţională.[3]

Clusteringul ierarhic descoperă clustere succesive, utilizând clusterele stabilite în prealabil, construind deci o ierarhie de clustere (producând o dendrogramă - diagramă arbore) şi nu doar o simplă partiţie a obiectelor.

[3] Florin Gorunescu, ***Data Mining. Concepte, Modele şi Tehnici***, Cluj-Napoca, Editura Albastră, 2006.

BDD-A19436 © 2010 Biblioteca Județeană Mureș
Provided by Diacronia.ro for IP 216.73.217.42 (2026-11-15 20:59:08 UTC)

Numărul clusterelor nu este cerut ca o condiţie input a algoritmului, în timp ce se poate utiliza o anumită condiţie de terminare a sa. Există trei tipuri de clustering ierarhic: aglomerativ (bottom-up), diviziv (top-down) şi conceptual, care constă într-un nod rădăcină gol, obiectele fiind adăugate unul câte unul (clustering incremental), utilizând clase deja existente, creând noi clase, combinând sau divizând clase existente.

Clusteringul neierarhic constă în partiţionarea mulţimii iniţiale a obiectelor în submulţimi (clustere) ce nu se suprapun, astfel încât fiecare obiect să aparţină exact unui cluster. Cei mai cunoscuţi algoritmi sunt cei aparţinând metodelor K-means şi Fuzzy K-means.

Avantajele şi dezavantajele în cazul clusterizării sunt strâns legate de algoritmul folosit.

În cazul algoritmului K-means, s-a constatat că acesta dă rezultate bune pe seturi de date foarte mari şi mai ales atunci când setul de date descrie o mulţime de puncte care formează grupe linear-separabile şi relativ îndepărtate una faţă de alta.

Între dezavantajele acestui algoritm amintim faptul că rezultatul final depinde de poziţiile iniţiale ale acelor centre de greutate K. În realitate, se recurge la calcularea mai multor variante cu modificarea poziţiilor iniţiale ale centrelor de greutate. Rezultatul final depinde foarte mult de metrica folosită la măsurarea distanţei şi de valoarea lui K.

În cazul clusterizării ierarhice de tip aglomerativ sau diviziv, apartenenţa obiectelor la clustere se face pe baza extragerii unor măsuri numerice ce exprimă similaritatea dintre obiecte. Aceste măsuri compară proprietăţi ale obiectelor, dar nu iau în considerare proprietăţi globale ce caracterizează clasele de obiecte. Clasele obţinute pot să nu aibă descrieri conceptuale clare şi, prin urmare, pot fi greu de interpretat.

Evaluarea calităţii partiţiei obţinute în urma aplicării unui algoritm de clusterizare este foarte importantă. Evaluarea trebuie să ia în considerare faptul că diferite metode conduc la clustere diferite.

Procedurile uzuale de evaluare includ:

- vizualizarea partiţiei (dendrograme, partiţii);
- analizarea indicatorilor de calitate.

Principalii coeficienţi ce exprimă calitatea operaţiei de clusterizare sunt:

BDD-A19436 © 2010 Biblioteca Județeană Mureș
Provided by Diacronia.ro for IP 216.73.217.42 (2026-11-15 20:59:08 UTC)

Coeficienţii de divizare (DC) - pentru fiecare obiect se calculează d(i) ca fiind raportul dintre diametrul ultimului cluster (în ordinea dată de algoritmul de divizare) la care a aparţinut obiectul înainte de a fi separat ca un singleton şi diametrul mulţimii totale de obiecte (clusterul iniţial).

Atunci:

$$\mathrm{DC} = \frac{1}{n}\sum d(i)$$

Coeficienţii de aglomerare (AC) - reprezintă indicii de calitate pentru clasificarea ascendentă. Pentru fiecare obiect *i* se calculează d(i) ca fiind raportul dintre disocierea primului cluster (în ordinea dată de algoritm), la care se ataşează obiectul şi diametrul mulţimii totale de obiecte (clusterul final).

$$\mathrm{AC} = \frac{1}{n}\sum (1 - d(i))$$

Literatura de specialitate prezintă şi alţi indici care exprimă calitatea clusterizării, dintre care amintim: *Indicele Dunn* (Dunn, 1974), *Indicele Davies-Bouldin* (Davies&Bouldin, 1979), *Indicele de siluetă* (Rousseeuw, 1987), *Indicele de precizie* (Topchy et al., 2003).

Maşini cu Suport Vectorial (Support Vector Machines - SVM). SVM sunt clasificatori ce folosesc o structură rafinată, care nu este necesar dependentă de dimensionalitatea spaţiului de intrare.

Ideea fundamentală care stă la baza clasificatorilor de tip SVM constă în găsirea unui plan optim ce poate separa vectorul spaţiu, astfel încât să evidenţieze cel mai bine membrii claselor diferite. Metoda a fost iniţiată de Vapnik (1995).

Conceptual, funcţionarea SVM-urilor se bazează pe următorii doi paşi:[4]

- aplicarea (neliniară) a spaţiului input într-un spaţiu de dimensiune mare - spaţiul caracteristicilor - spaţiu „ascuns," atât pentru input, cât şi pentru output;

[4] Daniel I. Morariu, ***Text Mining Methods Based on Support Vector Machine***, Bucureşti, Editura MatrixRom, 2008.

BDD-A19436 © 2010 Biblioteca Județeană Mureș
Provided by Diacronia.ro for IP 216.73.217.42 (2026-11-15 20:59:08 UTC)

- construirea unui hiperplan de separaţie pentru caracteristicile obţinute la pasul anterior.

În esenţă, SVM-urile mapează vectorul de intrare x pe un spaţiu asociat – z - cu mai multe dimensiuni decât x. În acest spaţiu asociat se va realiza separarea liniară printr-un hiperplan care separă un set de exemple pozitive de un set de exemple negative cu o limită maximă. Limita este definită de distanţa hiperplanului faţă de cele mai apropiate exemple pozitive şi negative. Problema de optimizare a SVM este găsirea unei suprafeţe de decizie care maximizează limita dintre punctele datelor din probleme de clasificare.

Ca *avantaje* ale acestui model amintim faptul că fundamentarea teoretică solidă conferă posibilitatea generalizării pe largi structuri de date.

Capacitatea mare de calcul necesară în procesare şi în proiectarea funcţiei nucleu (kernel) se numără printre *dezavantajele* acestui model.

3. Concluzii

Modelele matematice facilitează înţelegerea unui subiect şi îmbunătăţesc predicţia lui.

Existenţa unui anumit model matematic la baza unui software de bibliotecă determină tipul de management ce se va face în cadrul structurii info-documentare. Modelele matematice clasice (boolean, vectorial, probabilistic) determină existenţa unui management al datelor. Modelul boolean standard (practic, singurul folosit în soft-urile bibliotecilor româneşti) furnizează doar descrieri la nivel sintactic a datelor, fără nici o descriere formală a semanticii acestora. În acest context, datele sunt privite ca reprezentări cifrice sau letrice ale unor documente. Datele sunt uşor de captat şi organizat, uşor de transferat către maşini şi uşor de prelucrat.

Privind informaţiile ca un rezultat al unei prelucrări superioare a datelor, ca semnificaţie ce poate fi desprinsă dintr-un ansamblu de date, pe baza asociaţiilor dintre acestea, managementul informaţiilor nu poate fi furnizat prin simpla modelare - în sensul transpunerii într-un set de ecuaţii şi relaţii matematice - a intrărilor şi ieşirilor din sistem.

Lipsa unor semantici interpretabile de calculator necesită intervenţia umană pentru descoperirea şi compunerea datelor, ceea ce împiedică

BDD-A19436 © 2010 Biblioteca Județeană Mureș
Provided by Diacronia.ro for IP 216.73.217.42 (2026-11-15 20:59:08 UTC)

exploatarea fondului de documente în contexte complexe, în care automatizarea acestor procese e necesară.

Trecerea de la un management al datelor la un management al informaţiilor în cadrul unui software de bibliotecă se poate face doar prin dezvoltarea de aplicaţii specifice, care transformă datele în informaţii. Acest lucru se poate face prin punerea în relaţie a datelor la un nivel superior, şi anume la nivelul conceptelor la care se referă datele. Singurele metode matematice care implică acţiuni de punere în relaţie conceptuală a datelor sunt: Indexarea Semantică Latentă şi gruparea conceptuală (conceptual clustering).

Soluţia viabilă a unui software de management al informaţiilor este dată, în opinia noastră, de folosirea unui model matematic care transformă datele în informaţii prin intermediul conceptelor la care se referă datele.

BDD-A19436 © 2010 Biblioteca Județeană Mureș
Provided by Diacronia.ro for IP 216.73.217.42 (2026-11-15 20:59:08 UTC)