

## Optimizarea regăsirii informațiilor prin modelarea matematică

MIHAELA VOINICU

Biblioteca Județeană „Dinicu Golescu” Argeș

### 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 funcții ale bibliotecii este aceea de regăsire a informațiilor, 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 soluții în cadrul softurilor de bibliotecă.

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

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

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

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

Avantajele oferite de modelarea matematică sunt următoarele:

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

Prezentarea într-o formă agregată a modelelor matematice ce stau la baza regăsirii informațiilor se dovedește a fi un demers complicat, datorită varietății științelor sau domeniilor științelor ce tratează acest subiect (matematica, inteligența artificială, lingvistica computațională, știința informării etc.).

O scurtă trecere în revistă a principiilor ce stau la baza acestor modele, a avantajelor și 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 Informațiilor.

## 2. Modele matematice în informatica documentară

**Modelul boolean.** Este unul dintre primele și cele mai simple modele de reperare a informațiilor. Acest model identifică trei tipuri de relații de dependență, datorită operatorilor booleani AND, OR și NOT. Pe baza combinării descriptorilor cu ajutorul operatorilor booleani se urmărește suprapunerea exactă între document și interogare. Modelul boolean evaluează ce descriptori sunt prezenți sau absenți din document, obținând astfel o sumă de documente relevante (descriptori prezenți) și excluzând documentele irelevante (descriptorii lipsesc din document).

*Avantaje:* este un model robust, procesează rapid o interogare, scanează bine colecții mari de documente. Creșterea colecției de documente - cum este cazul bibliotecilor - nu reprezintă o problemă pentru acest model.

*Dezavantaje:* modelul boolean standard nu poate regăsi documentele pe baza relațiilor semantice sau conceptuale între descriptori; aceștia sunt tratați ca fiind independenți unul de celălalt, egali în importanță. Prin urmare, rezultatul unei interogări nu poate fi ordonat după relevanță, obținându-se liste de documente foarte mari și greu de exploatat.

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

Logica tradițională consideră că un obiect (în cazul nostru, un document) poate aparține, sau nu, unei mulțimi. În același 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 *parțial relevante*. În acest caz problema se rezumă la stabilirea transformării unor astfel de expresii în funcții de apartenență, asociate cu logica vagului.

Logica fuzzy permite o interpretare mai flexibilă a noțiunii de apartenență. Astfel, mai multe obiecte (documente) pot aparține unei mulțimi în grade diferite. Având în vedere aplicațiile 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 noțiunile fundamentale

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

Dacă fiecărui termen (descriptor) ce îi este atașat unui document în procesul de indexare și 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 obține liste de documente realizate pe baza operațiilor cu mulțimi fuzzy (reuniune, intersecție, inferență). În practică, decizia stabilirii importanței unui descriptor atașat unui document se dovedește a fi total subiectivă și 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 colecție. De obicei, acești descriptori sunt termeni care sunt extrași 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 semnificația lor în caracterizarea documentului. Aceste ponderi pot fi binare, indicând existența (1) sau nu (0) a termenilor în document.

O altă metodă de ponderare a termenilor, mult mai răspândită, este folosirea frecvenței de întâlnire a termenului în document sau un algoritm aparținând familiei Tf\*Idf. Frecvența de întâlnire a termenului este bazată pe statistica aparițiilor termenului în document și este cel mai simplu mod de a atribui ponderi unui termen. Tf\*Idf este o măsură folosită în colecțiile de documente care favorizează acei termeni care se regăsesc frecvent în documente relevante, dar sunt puțin frecvenți în colecție ca întreg. Această situație se întâlnește mai ales în cazul termenilor strict științifici, de foarte îngustă specializare. Tf reprezintă frecvența de întâlnire a termenului în document, iar Idf este inversul frecvenței de întâlnire a termenului în întreaga colecție.

$$\text{Idf} = \log (n_k/N)$$

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

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

În literatura de specialitate sunt menționate și alte încercări de ponderare a termenilor ce caracterizează un vector document. Astfel, se pot acorda ponderi diferențiate în funcție de locul în care se găsesc descriptorii în cadrul documentului (în titlu, în abstract, în conținut sau la concluzii) sau în funcție de modul în care sunt scriși (cu litere aldine sau italice), deoarece se presupune că numai termenii semnificanți sunt notați cu astfel de caractere.

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

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

**Modelul probabilistic.** Acest model se bazează pe calcularea probabilității ca un document  $d$ , ce aparține colecției de documente  $C$ , să fie relevant pentru utilizator.

La baza acestui model stă Algoritmul Naive Bayes care se folosește 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 aparțină unei categorii; adică se poate obține probabilitatea posterioară dacă se cunoaște probabilitatea anterioară. Estimarea acestor probabilități se face prin măsurarea frecvenței de apariție a cuvintelor într-un set de documente de antrenare. Această metodă se bazează pe simplificarea supozițiilor (independența reciprocă a atributelor ce caracterizează documentul) de unde vine și denumirea de Naive. Presupunerea că fiecare cuvânt sau descriptor al unui document este independent (are o apariție neafectată) de prezența sau absența oricărui alt cuvânt sau descriptor din document stă la baza formulării matematice a acestei teorii.

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

Probabilitatea ca un document  $d$  să aparțină clasei  $c$  se calculează ca:

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

unde  $P(t_i | c)$  este probabilitatea ca termenul  $t_i$  să aparțină unui document din clasa  $c$ .  $P(t_i | c)$  este interpretat ca o măsură a evidențierii a cât de mult contribuie  $t_i$  la desemnarea clasei  $c$ .  $P(c)$  este posibilitatea apriorică ca un document să aparțină clasei  $c$ .  $\{t_1, t_2, \dots, t_{n_d}\}$  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  $P(c)$  și  $P(t_i | c)$ , notăm cu  $\hat{P}(c)$  valoarea estimată din mulțimea de antrenament.

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

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

Astfel, estimăm:

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

unde  $T_{ct}$  reprezintă numărul de apariții 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:<sup>1</sup>

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

<sup>1</sup> C. V. Negoită, *Modelarea matematică în documentare*, București, Institutul Central de Documentare Tehnică, 1971.

- predictorii (variabilele independente ale modelului) - reprezentați de caracteristicile datelor ce urmează a fi clasificate, pe baza cărora se face clasificarea;
- mulțimea de antrenament (învățare) - care este reprezentată de setul de date care conține valori pentru cele două componente anterioare și care este utilizată pentru antrenarea modelului ca să recunoască clasa corespunzătoare, pe baza predictorilor disponibili;
- mulțimea de testare - conține date noi care vor trebui clasificate de modelul construit și, astfel, se va putea evalua acuratețea clasificării, adică performanța modelului.

Modelul probabilistic de regăsire a informațiilor relevante operează recursiv și necesită ca algoritmul din spatele metodei să fie inițializat cu parametri care, apoi, prin calcul iterativ, sunt îmbunătățiți, astfel încât să se obțină un scor ce desemnează relevanța probabilă.

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

Între *dezavantajele* acestei metode amintim: complexitatea metodei crește rapid; pe măsura creșterii colecției de documente, scanarea acestei colecții este din ce în ce mai greoaie; presupune unele metode simplificatoare, precum independența termenilor.

Luarea în considerare a unui istoric al interogărilor utilizatorului, în cadrul acestui model, poate îmbunătăți stabilirea parametrilor inițiali ai mulțimii de antrenament.

**Modelul lingvistic.** Din punct de vedere științific, limbajul natural (uman) constituie obiectul de cercetare a numeroase discipline și, în primul rând, al lingvisticii sau „științei limbii.” Același obiect de investigație interesează însă și filozofia, psiholingvistica, lingvistica matematică, lingvistica computațională, prelucrarea limbajului natural.

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

Această metodă folosește o matrice rară care conține pe coloane documentele în care se face căutarea, iar pe linii termenii (de obicei termeni trecuți prin procedura de stemmer - extragerea rădăcinii termenilor) conținuți în aceste documente. LSA-ul transformă această matrice a aparițiilor într-o relație dintre termeni și concepte și o relație între acele concepte și documente. Astfel, termenii și documentele sunt indirect legați 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 informațiilor a fost propusă de Deerwester la începutul anilor '90.

În cazul aplicării SVD, matricea inițială  $A$  este descompusă în produsul a trei matrice:<sup>2</sup>

$$A = T_{t \times n} S_{n \times n} (D_{d \times n})^T$$

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

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

$$\hat{A} = T_{t \times k} S_{k \times k} (D_{d \times k})^T$$

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 eliminați prin procesul de SVD. Comparând elementele matricei astfel reconstruită cu matricea inițială, se observă cum această metodă induce relații de similaritate, relații ce aproximează judecata umană asupra înțelesului și similarității 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 colecție cu un vocabular eterogen,

---

<sup>2</sup> Thomas K Landauer & al, *Handbook of Latent Semantic Analysis*, New Jersey, Lawrence Erlbaum Associates Inc. Publishers, 2007.

unde documentele pot folosi termeni diferiți pentru a face referire la același subiect.

Între *dezavantaje* amintim: apariția simultană a termenilor poate induce false rezultate și o precizie mai scăzută; în colecții de documente cu un vocabular omogen, rezultatele acestei metode sunt mai puțin folositoare.

**Modelul bazat pe clustere.** Clusterizarea este una dintre tehnicile fundamentale din domeniul analizei datelor care organizează un set de obiecte dintr-un spațiu multidimensional, în grupuri coezive, numite clustere. Tehnicile de clustering (grupare spațială) reprezintă tehnici speciale de aranjare a datelor de intrare pe baza dispunerii spațiale a vectorilor corespunzători. Pentru a analiza asemănarea/deosebirea dintre elementele unei mulțimi, 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 funcție de situație, care reprezintă poziționarea spațială a proprietăților 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 aparțin aceleiași grupe dacă aceasta definește un concept comun tuturor elementelor. Altfel spus, elementele sunt grupate în conformitate cu potrivirea la conceptele descriptive și nu în conformitate cu măsurile de similaritate.

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

Metodologia clustering are două abordări distincte: clusterizarea ierarhică și clusterizarea neierarhică/partițională.<sup>3</sup>

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

---

<sup>3</sup> Florin Gorunescu, *Data Mining. Concepte, Modele și Tehnici*, Cluj-Napoca, Editura Albastră, 2006.

Numărul clusterelor nu este cerut ca o condiție input a algoritmului, în timp ce se poate utiliza o anumită condiție de terminare a sa. Există trei tipuri de clustering ierarhic: aglomerativ (bottom-up), diviziv (top-down) și 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 partiționarea mulțimii inițiale a obiectelor în submulțimi (cluster) ce nu se suprapun, astfel încât fiecare obiect să aparțină exact unui cluster. Cei mai cunoscuți algoritmi sunt cei aparținând metodelor K-means și Fuzzy K-means.

*Avantajele și 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 și mai ales atunci când setul de date descrie o mulțime de puncte care formează grupe linear-separabile și relativ îndepărtate una față de alta.

Între dezavantajele acestui algoritm amintim faptul că rezultatul final depinde de pozițiile inițiale ale acelor centre de greutate K. În realitate, se recurge la calcularea mai multor variante cu modificarea pozițiilor inițiale ale centrelor de greutate. Rezultatul final depinde foarte mult de metrica folosită la măsurarea distanței și de valoarea lui K.

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

Evaluarea calității partiției obținute în urma aplicării unui algoritm de clusterizare este foarte importantă. Evaluarea trebuie să ia în considerare faptul că diferite metode conduc la cluster diferite.

Procedurile uzuale de evaluare includ:

- vizualizarea partiției (dendrograme, partiții);
- analizarea indicatorilor de calitate.

Principalii coeficienți ce exprimă calitatea operației de clusterizare sunt:

*Coeficienții 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 aparținut obiectul înainte de a fi separat ca un singleton și diametrul mulțimii totale de obiecte (clusterul inițial).

Atunci:

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

*Coeficienții 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 atașează obiectul și diametrul mulțimii totale de obiecte (clusterul final).

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

Literatura de specialitate prezintă și alți 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).

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

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

Conceptual, funcționarea SVM-urilor se bazează pe următorii doi pași:<sup>4</sup>

- aplicarea (neliniară) a spațiului input într-un spațiu de dimensiune mare - spațiul caracteristicilor - spațiu „ascuns,” atât pentru input, cât și pentru output;

---

<sup>4</sup> Daniel I. Morariu, *Text Mining Methods Based on Support Vector Machine*, București, Editura MatrixRom, 2008.

- construirea unui hiperplan de separație pentru caracteristicile obținute la pasul anterior.

În esență, SVM-urile mapează vectorul de intrare  $x$  pe un spațiu asociat  $z$  - cu mai multe dimensiuni decât  $x$ . În acest spațiu 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 distanța hiperplanului față de cele mai apropiate exemple pozitive și negative. Problema de optimizare a SVM este găsirea unei suprafețe 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 și în proiectarea funcției nucleu (kernel) se numără printre *dezavantajele* acestui model.

### 3. Concluzii

Modelele matematice facilitează înțelegerea unui subiect și îmbunătățesc predicția lui.

Existența 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 existența unui management al datelor. Modelul boolean standard (practic, singurul folosit în soft-urile bibliotecilor românești) 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 ușor de captat și organizat, ușor de transferat către mașini și ușor de prelucrat.

Privind informațiile ca un rezultat al unei prelucrări superioare a datelor, ca semnificație ce poate fi desprinsă dintr-un ansamblu de date, pe baza asociațiilor dintre acestea, managementul informațiilor nu poate fi furnizat prin simpla modelare - în sensul transpunerii într-un set de ecuații și relații matematice - a intrărilor și ieșirilor din sistem.

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

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 informațiilor în cadrul unui software de bibliotecă se poate face doar prin dezvoltarea de aplicații specifice, care transformă datele în informații. Acest lucru se poate face prin punerea în relație a datelor la un nivel superior, și anume la nivelul conceptelor la care se referă datele. Singurele metode matematice care implică acțiuni de punere în relație conceptuală a datelor sunt: Indexarea Semantică Latentă și gruparea conceptuală (conceptual clustering).

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