ANNALS of the ORADEA UNIVERSITY.

Fascicle of Management and Technological Engineering

Rolul bibliotecilor in organizarea sistemului informational al universitatii

Rodica Volovici

Universitatea “Lucian Blaga” din Sibiu
Bd. Victoriei Nr. 10
e-mail:rodica.volovici@ulbsibiu.ro

Abstract: Information technology and communication technology have fundamentally changed our life. They have not missed libraries. Since the discovery of printing it gathers and contains written and printed paper and makes it accessible under the form of books.

În cadrul relatiei biblioteca-cercetare, legaturile simple pâna nu demult, capata în noua societate informationala o complexitate si o dinamica neasteptate, care le modifica termenii si desfasurarile. Se stie ca bibliotecile au fost si sunt principala baza de referinta a tuturor dezbaterilor, conjuncturilor si risipelor gândului. Aceasta motiveaza cu o temeinicie greu de contestat validitatea relatiei biblioteca-cercetare. Si totusi, pentru ca biblioteca sa devina cu adevarat o componenta sine qua non a cercetarii de calitate în conditiile exploziei de cunoastere determinata strâns de dezvoltarea tehnicii informatiei, se impune o reorganizare a lucrului cu documentele si a gestiunii bibliotecilor.

Cu ajutorul noilor tehnici ale prelucrarii informationale, activitatea de informare-documentare poate fi adusa mai aproape si mai repede catre beneficiarii bibliotecilor, apropiind fazele finale ale muncii creatoare. Calitatea informarii documentare realizata de o biblioteca poate sa creasca substantial prin îmbunatatirea unor indicatori de referinta între care figureaza cu necesitate:

- numarul de directii si teme prioritare acoperite cu informatii,
- cantitatea de informatie relevanta oferita beneficiarilor,
- distanta în timp între aparitia documentelor si furnizarea informatiei.

Abordarea problemelor referitoare la eficientizarea activitatii bibliotecilor impune oricarui analist o relatie strânsa cu teoria biblioteconomica si cu stiinta informarii. Biblioteca poate fi privita ca un ansamblu cu functii si activitati dependente de structuri, relatii, dotari, metode, de organizarea si calitatea acestor componente si, nu în ultimul rând, de competenta. Gasim aici o colectie de entitati care sunt în relatie unele cu altele si cu mediul, astfel încât ele formeaza un întreg, un sistem complex.

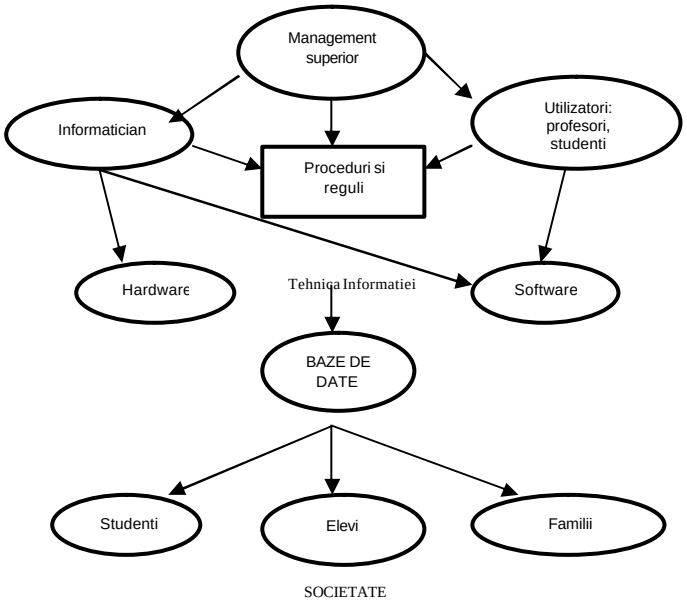
Un concept cheie care rezulta din definitia sistemului este ca elementele componente ale acestuia luate fiecare separat, nu prezinta si nu reprezinta esenta sistemului. Este necesar sa consideram sistemul ca întreg pentru a-l întelege. Deci întregul este mai mare decât suma partilor sale; aceasta proprietate este cunoscuta sub numele de sinergie. Aceasta este o proprietate evidenta a sistemelor organizationale în general, si în particular a sistemelor informationale.

O prima schita a sistemului informational universitar, nu trebuie sa prezinte toate conexiunile în complexitatea lor, ci numai elementele constitutive cele mai importante si cum interactioneaza aceste elemente. De asemenea, trebuie avut în vedere si cum interactioneaza sistemul cu alte sisteme ("furnizorii" si "clientii").

Optiunea pentru modernizare, pentru introducerea celor mai noi tehnici de achizitie, prelucrare, stocare si dimensionare a documentelor si informatiilor este si aici, ca în orice domeniu de activitate, o conditie hotarâtoare a cresterii eficientei. Transferul de informatie trebuie sa se realizeze în conditii si cu mijioace competitive cu cele ale structurilor care beneficiaza de acest proces, daca nu chiar în situatii privilegiate, dat fiind rolul propulsiv al informatiei stiintifice în toate sferele de cunoastere si actiune umana. Daca unul din factorii de baza ai eficientei este viteza de regasire a informatiilor, nici un sistem traditional nu poate oferi adrese si combinatii într-un timp mai scurt. Dar impactul informatizarii bibliotecilor asupra vietii economice si sociale este mult mai puternic si mai adânc.

Oricare ar fi determinarile materiale ale eficientei în activitatea bibliotecilor, principala conditie a unor rezultate remarcabile o constituie calitatea specialistilor antrenati. Însusirea obiectivelor economico-sociale, prin traducerea cerintelor în limbajul si tehnicile informarii documentare, organizarea achizitiei, prelucrarii, stocarii si diseminarii informatiei si documentelor, sustinerea unui dialog fertil cu beneficiarii, nu pot fi decât rodul unei participari profesionale, politice si cetatenesti marcata de competenta si clarviziune.

A vedea este aproape similar cu a citi. Este atunci firesc primul pas facut de programatori în directia aceasta, si anume, sa învete calculatorul sa recunoasca literele. Tentativa este cunoscuta sub numele de OCR, acronim pentru Optical Character Recognition, adica recunoasterea optica a caracterelor. Avantajul unui astfel de soft este evident atunci când se doreste stocarea documentelor într-o forma electronica (text ASCII). Marea majoritate a sistemelor OCR încep cu o imagine bit-map culeasa via scanner sau fax-modem si primul pas consta din identificarea pe pagina a blocurilor de text, pe baza unor caracteristici tipografice, cum ar fi alinierea la stânga sau la dreapta. Apoi aceste blocuri sunt descompuse în semne individuale, care deseori corespund literelor.



Algoritmul de recunoastere încearca în urmatoarea etapa sa "ghiceasca" cât mai bine semnificatia fiecarui semn, dupa care este construita întreaga pagina cu formatul corespunzator. Cel mai bun sistem OCR poate realiza achizitii cu o acuratete de peste 99% pentru paginile tiparite si scrise cu fonturi obisnuite. Cu toate ca acest procent pare aproape perfect, nivelul de eroare este destul de ridicat daca ne gândim ca pe o pagina standard sunt aproape 1500 de caractere. În acest caz, chiar si un procent de corectitudine de 99,9% genereaza una sau doua erori pe pagina necesitând interventia omului pentru corectare si asigurarea fidelitatii la citire.

În viata de zi cu zi ne întâlnim rareori cu situatii ideale: text "curat", font standard, litere tiparite complet, astfel ca un "o" sa nu semene cu un "c". Imaginile murdare sunt o problema deoarece chiar si o pata mica poate acoperi o parte importanta din litera facând un "m" sa semene cu un "n". Daca documentul este neclar, daca contururile literelor sunt marite sau lipite unele de altele, pot sa apara erori de identificare a caracterelor, deoarece sistemele OCR considera orice semn continuu ca un singur caracter. Probleme pot sa apara chiar si daca paginile sunt "curate". Dezvoltarea algoritmilor care pot recunoaste caracterele, în ciuda acestor probleme, trebuie sa asigure un echilibru între flexibilitate si acuratete. Daca soft-ul nu este destul de flexibil va da eroare daca întâlneste cea mai mica diferenta între corpurile de litere si, pe de alta parte, o mare flexibilitate poate conduce la confundarea a doua caractere asemanatoare cum ar fi "b" si "h".

Rezolvarile tehnice gasite în ultima vreme sunt foarte bune daca textul este "curat" si forma literelor identica în întregul text.

Softurile OCR lucreaza, de regula, pornind de la imaginea unei pagini "furnizate" de un scanner sau care este preluata de la un fax-modem, deci ele pot conversa direct cu scannerele, iar unele se pot atasa aplicatiilor astfel încât se poate translata o pagina de text direct în editorul de texte.

Primul pas al procesului OCR este "spargerea" imaginii paginii în blocuri de text, grafice, imagini, tabele. Exista algoritmi specifici pentru identificarea naturii fiecarui bloc. În general rândurile sunt grupate în paragrafe utilizând criterii cum ar fi: alinierea, lungimea, corpul de litera, spatierea. Pentru a se face distinctia între grafica si text, de regula, se utilizeaza algoritmi bazati pe analiza diferentelor statistice între text si grafica, precum: ritm regulat la text fata de lipsa unei periodicitati la grafica.

Dupa reconstruirea paginii, semnele individuale sunt translatate în text ASCII. Exista mai multe modalitati de a face acest lucru, dar în mare exista doua categorii: metode bazate pe sabloane (template-based) si metode bazate pe trasaturi caracteristice (feature-based). Dintre sistemele OCR bazate pe sabloane existente acum pe piata amintim sistemul Expert Vision, care recunoaste pata de cerneala comparând-o cu sablonul cu care se aseamana cel mai mult. Un exemplu elocvent de soft care utilizeaza cealalta metoda de recunoastere, cea bazata pe trasaturi specifice, este OmniPageProfessional (Caere) care prezinta avantajul ca nu trebuie "reglat" pentru anumite corpuri de litera. Softul contine aproximativ 100 de "sisteme expert", în esenta 100 de algoritmi, pentru identificarea unor caractere de la "A" la "Z" si apoi de la "a" la "z", la care se adauga numerele si semnele de punctuatie. Acest tip de soft încearca sa surprinda esenta fiecarei litere într-un algoritm.

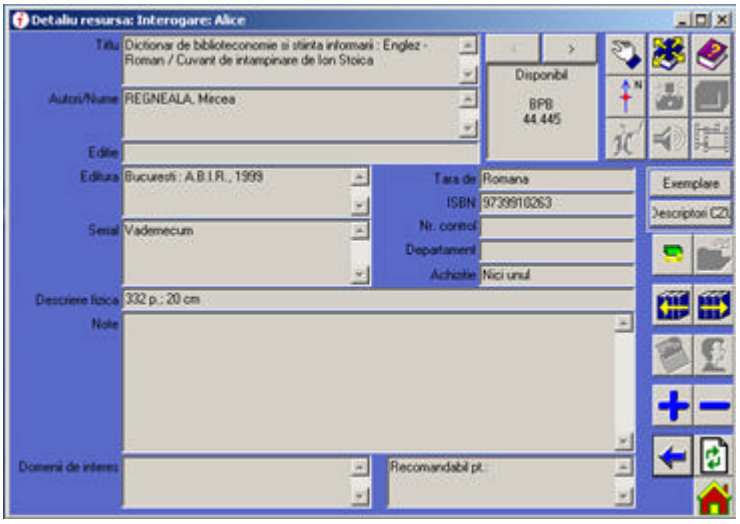
Unele dintre produsele software specializate în recunoasterea caracterelor sunt orientate pe context, ele contin o serie informatii cu privire la regulile de scriere existente si la formularile cele mai uzuale dintr-o limba sau un grup de limbi. Si asta, deoarece majoritatea softurilor OCR sunt dedicate lingvistic. Indiferent de tehnologia de recunoastere utilizata, semnele diacritice, sau în alte cazuri, chiar forma diferita a caracterelor pot fi motive de reducere drastica a procentului de recunoastere. Este un lucru obisnuit ca acelasi soft OCR sa fie utilizat pentru o singura limba sau pentru un grup de limbi. De pilda Recognita

(Recognita Corporation) este implementat în mai multe versiuni: numai pentru engleza, pentru limbi vorbite pe continentele Americane (engleza, franceza, spaniola, portugheza) si o versiune internationala care ofera posibilitatea de a alege din 22 de limbi, una dintre ele find si româna.

Tehnologiile de ultima ora, cum sunt de pilda cele bazate pe modelul retelelor neuronale, ofera si posibilitatea autoinstruirii. Un exemplu este produsul Five-Star-Read (Epson). Softul poseda functii speciale pentru a învata cum anume trebuie interpretat un anumit semn, dezvoltate din cercetarile privind recunoasterea scrisului de mâna. Daca textul care trebuie recunoscut este degradat si algoritmul de recunoastere da în prima instanta gres, majoritatea softurilor încearca un proces de îmbunatatire a imaginii caracterelor, de pilda adaugând un pixel lipsa când este posibila refacerea unei linii mai lungi. Dar, foarte frecvent si aceste procedee dau gres. Si atunci, cel mai bun OCR ramâne în continuare omul. Verificarea corectitudinii se dovedeste a fi cel mai mare consumator de timp în întregul proces OCR.

Echipa noastra si-a propus sa studieze acuratetea cu care sistemele OCR existente acum pe piata sunt capabile sa recunoasca caracterele de pe vechile fise de carte. Se vor analiza comparativ performantele obtinute în recunoasterea fiecarui caracter si se vor identifica situatiile în care sitemele existente sunt suficient de perfomante pentru a putea fi folosite pentru ducerea la bun sfârsit a obiectivului propus, si, se vor identifica si situatiile în care performantele sunt sub un nivel minim acceptabil. Pentru aceste ultime situatii, se vor dezvolta noi sisteme OCR implementate fie prin metode clasice fie folosind retele neuronale, astfel încât în final sa se obtina un sistem OCR deosebit de performant.

În concluzie, se va implementa un sistem OCR dedicat, deosebit de performant care sa poata fi folosit pentru conversia fiselor vechi de carte, existente în format dactilografiat, în fise electronice de carte care sa poata fi încorporate în baza de date a Bibliotecii Universitare Sibiu.



Bibliografie:

1. Thompson, Ronald; Cats-Baril, William: Information Technology and Management, 2nd edition, Boston: McGraw-Hill, 2003
2. Volovici, R. M.: Cercetari asupra cresterii fiabilitatii sistemului informatic si al informatiei, referat de doctorat, ULB Sibiu, Sibiu, 2003