STUDIA UNIVERSITATIS MOLDAVIAE, 2015, nr.2(82)
Seria "Ştiinţe exacte şi economice" ISSN 1857-2073 ISSN online 2345-1033 p.130-139

NOI GENERAŢII DE SISTEME DE RECOMANDARE ŞI ALGORITMI PENTRU IMPLEMENTAREA LOR ÎN COMERŢUL ELECTRONIC

Gheorghe MIŞCOI, Elena BĂDĂRĂU*

Universitatea Liberă Internaţională din Moldova,
**Institutul de Relaţii Internaţionale din Moldova*

În articol a fost efectuată o analiză a problemelor de bază ale sistemelor de recomandare, a metodelor de elaborare a noi sisteme de recomandare, mai adecvate proceselor reale, şi a algoritmilor efectivi de realizare a lor.

Cuvinte-cheie: *sisteme de recomandare, procese semimarkoviene, priorităţi, flux Poisson, modelare matematică, algoritm Bayes, algoritmul aproximărilor succesive.*

NEXT-GENERATION SYSTEMS OF RECOMMENDATION AND ALGORITHMS FOR IMPLEMENTATION IN E-COMMERCE

In the article was given an analysis of the basic problems of recommendation systems, methods of developing new systems for recommendation, more adequately to real processes and effective algorithms of their realization.

Keywords: *recommendation systems, semimarkovien processes, priorities, Poisson flow, mathematical modeling, Bayes algorithm, algorithm of successive approximations.*

Introducere: Internetul contemporan oferă utilizatorului o cantitate enormă de informaţie diversificată, în care este tot mai dificil de a te orienta şi naviga, iar sistemele clasice de căutare şi sistematizare a informaţiei nu mai verifică necesităţile utilizatorului. Ca urmare, au apărut aşa-numitele Sisteme de Recomandare (SR), orientate spre furnizarea de informaţii la maxim posibil verificând interesele utilizatorului. SR reprezintă soft-uri prin intermediul cărora se încearcă de a elabora pronosticuri referitor la bunurile şi/sau serviciile (cărţi, filme, algoritmi, hoteluri, restaurante etc.) pe care le-ar dori utilizatorul având anumite informaţii despre profilul lui [3, 8].

SR au devenit obiectul atenţiei ştiinţifice începând cu mijlocul anilor '90 ai secolului XX, de la primele lucrări cu privire la filtraţia colaborativă. În ultimul deceniu SR au fost dezvoltate atât sub aspect teoretic, cât şi aplicativ, dar în acest domeniu rămân multe probleme, a căror soluţionare promite vaste aplicări practice ce ar permite utilizatorului să se lupte cu volumul mare de informaţie şi să-l înarmeze cu instrumente de elaborare a recomandărilor personificate. Exemple de aplicaţii ale SR pot servi sistemele de recomandare a cărţilor, CD-urilor şi altor mărfuri pe Amazon.com, a filmelor pe Netflix şi MovieLens, a noutăţilor pe VERSIFI Technologies etc.

Rezultatele cercetării şi analiza: Perfecţionarea în continuare a SR este necesară pentru o utilizare mai vastă şi mai efectivă în situaţii reale pentru elaborarea de recomandări ce se referă, de exemplu, la călătorii, servicii financiare pentru investori, mărfuri în magazine procurate prin intermediul cardului „inteligent" al consumatorului, diferite servicii. Devin necesare metode noi pentru urmărirea comportamentului consumatorului şi informaţia propusă lui, modele performante de elaborare a recomandărilor ce includ informaţia referitor la contextul de procurare, utilizarea unui sistem multifactorial de estimare, reguli speciale de prioritate, modele de comportament virtual, precum şi folosirea sau adoptarea diferitelor compartimente ale matematicii contemporane, ceea ce ar permite elaborarea unor SR mai flexibile şi mai puţin sofisticate.

În acest context se înscrie şi ipoteza despre evoluţia pieţelor. Nu demult era important pe cât e de calificativă marfa. Pieţele mai complexe consideră important nu atât „valorile reale", cât cum sunt ele percepute de consumatori. Acum începe o perioadă a „pieţelor noi", când primordial devine nu „valoarea" produsului sau chiar „percepţia" lui, ci ce este cunoscut despre percepţia acestui bun de alţi consumatori. Motorul pieţelor moderne constă în abilitarea consumatorului de a adera la consumul colectiv sau a se distanţa de el; devine important câţi oameni şi cum percep produsul dat. Astfel, elaborarea şi dezvoltarea SR, capabile să dea recomandări în „timp real", să efectueze pronosticuri, devine stringentă pe piaţa tehnologiilor informaţionale.

Pronosticul în SR se elaborează în baza datelor despre utilizator, care se obţin prin metode explicite (cereri de evaluare a obiectului de către consumator după o scară diferenţială, cereri de aranjare a unui grup

de bunuri de la cel mai bun la cel mai rău, prezentarea consumatorului a două obiecte cu întrebarea care este mai bun, elaborarea unei liste de preferinţe) şi implicite (urmărirea vizualizării consumatorului internet-magazinelor sau a diferitelor baze de date, înregistrarea comportamentului consumatorului on-line, conţinutul calculatorului etc.). În rezultatul datelor colectate SR determină un set de recomandări pentru utilizatorul concret.

Formal, problema elaborării recomandărilor poate fi reprezentată astfel: fie **C** – un grup (set) de consumatori, **S** – un grup (set) de bunuri propuse. Grupul **S** poate fi foarte voluminos, atingând în unele domenii sute de mii sau milioane de unităţi de bunuri (cărţi, CD etc.), grupul **C** la fel poate fi extrem de mare. Fie **U** – funcţia de utilitate ce descrie utilitatea bunului **s** pentru **c**, $U : C \times S \to R$, unde R – numărul de bunuri comandate. Astfel, pentru fiecare consumator $c \in C$ dorim să alegem aşa un bun $s \in S$ care ar fi cât mai util consumatorului.

În SR utilitatea unui bun este determinată pe cât i-a plăcut unui consumator acest bun. În funcţie de aplicaţie, utilitatea **u** poate fi determinată de consumator (de regulă, în sistemele bazate pe evaluările consumatorului) sau poate fi calculată automat, similar funcţiei de utilitate, bazate pe beneficiu.

Fiecare membru din mulţimea **C** poate fi descris prin intermediul unui profil individual, care conţine diverse caracteristici ale lui: vârsta, sexul, venitul, starea familială etc. În cel mai simplu caz profilul poate conţine un singur parametru – **ID**–consumatorului. Sau, de exemplu, în SR a filmelor, unde **S** – totalitatea filmelor, fiecare film poate fi reprezentat nu doar prin **ID**-ul său, dar şi prin denumire, gen, regizor sau actorii principali.

Cea mai importantă problemă a SR constă în faptul că utilitatea **u**, de regulă, nu este determinată pe întregul set **CxS**. În SR utilitatea este reprezentată, de regulă, prin puncte şi iniţial definită doar pentru bunurile deja apreciate de consumator.

Extrapolarea de la estimările cunoscute la cele necunoscute are loc, de regulă, astfel: 1) prin alegerea unor reguli euristice ce determină funcţia de utilitate şi justifică comportamentul ei euristic *sau* 2) determinarea funcţiei de utilitate ce optimizează comportamentul parametrilor ei daţi (abaterea medie pătratică).

Cum numai estimările necunoscute au fost analizate, consumatorul primeşte ca recomandare bunurile **c** cu cele mai înalte aprecieri [8]. SR pot determina sau totalitatea **N** de bunuri (cele mai potrivite) pentru consumator sau totalitatea consumatorilor relevanţi bunului dat. Estimările noi pentru bunurile neestimate încă pot fi făcute prin diferite metode, de exemplu: euristice, teoria aproximării, luarea deciziilor.

SR diferă în abordarea analizei estimărilor. Clasificarea modernă a SR le împarte în următoarele categorii în funcţie de modul cum sunt făcute recomandările:

- ***Recomandări de content:*** consumatorul primeşte recomandări de bunuri similare cu cele alese anterior;
- ***Recomandări colaborative:*** consumatorul primeşte recomandări de bunuri selectate anterior de persoane cu gusturi şi cerinţe similare;
- ***Recomandări ce combină cele două metode anterioare.***

Metodici de content. În SR de content concluzia referitor la utilitatea **u(c, s)** se ia reieşind din utilitatea **u(c, s_i)**, atribuită de consumatorul **c** bunului $s_i \in S$ similar bunului **s**. Mai formal, fie **Content (s)** profilul bunului **s**, adică totalitatea proprietăţilor bunului **s**. Acest set în continuare se foloseşte pentru a analiza utilitatea bunului pentru consumator. Cel mai frecvent, SR se utilizează pentru a recomanda bunuri ce conţin informaţie textuală. Contentul acestor sisteme se descrie folosind cuvinte-cheie.

„Importanţa" sau („informativitatea") cuvântului $\mathbf{k_i}$ în documentul $\mathbf{d_j}$ se determină prin ponderea lui, $\mathbf{w_{ij}}$, care poate fi determinată prin mai multe metode. Una din cele mai cunoscute metode de măsurare a ponderii cuvintelor în sistemele de căutare este aşa-numita metoda frecvenţei directe/indirecte, a cărei esenţă se reduce la: fie **N** – cantitatea de documente ce pot fi recomandate utilizatorului, cuvântul-cheie $\mathbf{k_j}$ se întâlneşte în $\mathbf{n_i}$ documente; în plus, $\mathbf{f_{ij}}$ – va indica de câte ori cuvântul-cheie $\mathbf{k_i}$ se întâlneşte în documentul $\mathbf{d_j}$.

Atunci $\mathbf{TF_{i,j}}$ – frecvenţa cuvântului-cheie $\mathbf{k_i}$ în documentul $\mathbf{d_j}$, se determină prin formula

$$TF_{i,j} = \frac{f_{i,j}}{\max\limits_{z} f_{z,j}}, \qquad (1)$$

unde maximul se ia după toate frecvenţele $\mathbf{f_{z,j}}$ ale tuturor cuvintelor-cheie $\mathbf{k_z}$, ce se întâlnesc în domeniul $\mathbf{d_j}$. Cu toate acestea, cuvintele-cheie ce se întâlnesc frecvent nu ajută suficient pentru distingerea unui document

relevant de unul nerelevant. De aceea, măsurarea frecvenţei inverse a cuvântului (**IDF$_i$**) este adesea utilizată de rând cu **TF$_{i,j}$**. Frecvenţa inversă a cuvântului-cheie **k$_i$** se defineşte de obicei prin

$$IDF_i = \log \frac{N}{n_i}, \tag{2}$$

ponderea cuvântului-cheie **k$_i$** în documentul **d$_j$** – prin

$$w_{i,j} = TF_{i,j} \cdot IDF_i, \tag{3}$$

iar contentul documentului **d$_j$** – prin Content (d_j)=(w_{1j},w_{2j},…w_{kj}).

Fie *Content Profil (c)* – profilul consumatorului **c** ce conţine informaţia despre gusturile şi preferinţele acestui consumator. Aşa tip de profiluri se creează ca rezultat al analizei contentului (proprietăţilor) bunurilor deja estimate de consumatori şi se bazează pe analiza cuvintelor-cheie. De exemplu, *Contentul Profilului (c)* poate fi definit ca un vector de liste (w_{c1},w_{c2},…,w_{ck}), unde fiecare listă w_{ci} determină importanţa cuvântului-cheie k_i şi poate fi calculată pe baza vectorilor contentului estimat individual.

În sistemele de content funcţia de utilitate, de regulă, se defineşte ca

$$u(c,s) = punctaj\,(contentul\ \ profilului\,(c), Content\,(s)) \ . \tag{4}$$

În utilizarea sistemului de căutare a informaţiei de mai sus pentru recomandări de web-sait-uri, URL, mesaje ale reţelei Usenet, Content Profil (c) pentru consumatorul **c** şi Content (s) a documentului **s** pot fi reprezentaţi ca vectori ai frecvenţei directe/ indirecte a ponderii cuvintelor-cheie. În plus, în literatura axată pe problemele de găsire a informaţiei, funcţia de utilitate **u(c,s)** de obicei se reprezintă ca un produs al unor algoritmi euristici, exprimaţi de vectorii $\overrightarrow{w_c}$ şi $\overrightarrow{w_s}$,

$$u(c,s) = \frac{\overrightarrow{w_c} \cdot \overrightarrow{w_s}}{\left|\overrightarrow{w_c}\right| \cdot \left|\overrightarrow{w_s}\right|} = \frac{\sum_{i=1}^{k} w_{i,c} \cdot w_{i,s}}{\sqrt{\sum_{i=1}^{k} w_{i,c}^2} \sqrt{\sum_{i=1}^{k} w_{i,s}^2}}, \tag{5}$$

unde k – numărul general de cuvinte-cheie în sistem. De exemplu, dacă consumatorul **c** citeşte un număr mare de articole on-line pe bioinformatică, SR de content vor putea recomanda acestui utilizator şi alte articole din acest domeniu, deoarece aceste articole conţin mai mulţi termeni speciali (ghenom, secvenţiere, proteomică etc.) decât în articole cu altă tematică şi, prin urmare, Content Profil (c) descris de vectorul $\overrightarrow{w_c}$ va reprezenta aşa termeni **k$_i$** cu o pondere **w$_{ic}$** mai mare.

În afară de metodele euristice tradiţionale, bazate pe principiile de căutare a informaţiei, există şi alte metode-content. Aceste metode sunt caracterizate prin aceea că ele au la bază datele anterioare obţinute cu ajutorul analizei statistice sau învăţare automată. De exemplu, pe totalitatea paginilor web, estimate de consumator ca „utile" sau „inutile", se defineşte clasificatorul Bayes pentru a clasifica paginile neestimate. Acest clasificator trebuie să prezică probabilitatea că pagina **p$_j$** aparţine clasei **C$_i$** (adică, este importantă sau nu) reieşind din cuvintele-cheie $k_{1,j}$, $k_{2,j}$, …, $k_{n,j}$ de pe această pagină:

$$P(C_i/\ k_{1,j}\ \&\ k_{2,j}\ \&\ldots\&\ k_{n,j}) \tag{6}$$

În [2], autorii reiese din presupunerea că toate cuvintele-cheie sunt independente şi, prin urmare, probabilitatea (6) este proporţională cu

$$P(C_i)\prod_x P(k_{x,j}|C_i) \tag{7}$$

În plus, probabilităţile $P(k_{x,j}/\ C_i)$ şi $P(C_i)$ pot fi estimate reieşind din datele deja existente. În aşa caz, pentru fiecare pagină **p$_j$** probabilitatea (6) se calculează pentru fiecare clasă **C$_i$** şi pagina **p$_j$** se atribuie la clasa C_i cu cea mai mare probabilitate.

Metodele colaborative. SR colaborative (sau sistemele de filtrare colaborativă) încearcă să prezică utilitatea bunului pentru un consumator aparte, reieşind din estimările făcute de alţi consumatori. Mai formal, utilitatea **u(c,s)** a bunului **s** pentru consumatorul **c** se estimează prin utilităţile **u(c$_j$,s)** atribuite bunului **s** de acei consumatori **c$_j$**, care „seamănă" cu consumatorul **c**.

Conform [11], algoritmii pentru filtraţia colaborativă pot fi divizaţi în două clase mari: amnestici (**memory-based**) şi de model (**model-based**).

Algoritmii amnestici, bazându-se pe ipoteze euristice, fac pronosticuri referitor la estimările consumatorului reieşind din toate estimările precedente date de acest consumator.

Valoarea estimării necunoscute $\mathbf{r_{c,s}}$ pentru consumatorul **c** şi bunul **s** se calculează, de regulă, din totalitatea estimărilor bunului **s** date de alţi consumatori. De exemplu:

$$r_{c,s} = \frac{1}{N}\sum_{c' \in C^*} r_{c',s} \tag{8}$$

$$r_{c,s} = k\sum_{c' \in C^*} sim(c,c') \cdot r_{c',s} \tag{9}$$

$$r_{c,s} = \bar{r}_c + k\sum_{c' \in C^*} sim(c,c') \cdot (r_{c',s} - \bar{r}_c), \tag{10}$$

unde: C^* – totalitatea a N consumatori cu gusturi similare consumatorului **c**, care au estimat deja bunul **s**;

k – coeficient de normare. De regulă,

$$k = \frac{1}{\sum_{c' \in C^*} |sim(c,c')|}.$$

Estimarea medie $\bar{r}_c$ a consumatorului **c** se defineşte ca

$$r_c = \frac{\sum_{s \in S_c} r_{c,s}}{|s_c|}; \tag{11}$$

sim(c,c') – măsură liniară a similitudinii consumatorilor **c** şi **c'**.

În cel mai simplu caz, însumarea poate fi o medie simplă (8). Totodată, de obicei, pentru însumare se ia suma ponderată (9). Măsura liniară a similitudinii de regulă serveşte ca măsură de distanţă dintre **c** şi **c'** şi se utilizează pentru măsurarea ponderii, adică cu cât sunt mai asemănători în gusturi **c** şi **c'** cu atât mai mare va fi ponderea estimării $\mathbf{r_{c',s}}$ în pronosticul $\mathbf{r_{c,s}}$. Trebuie de reţinut că *sim(x,u)* reprezintă un artefact euristic, necesar pentru a determina totalitatea „celor mai apropiaţi clienţi" pentru fiecare consumator.

În cazul aplicării (9) trebuie să se ia în considerare faptul că diferiţi consumatori pot folosi diferite scări de evaluare. Pentru a ocoli această restricţie putem utiliza suma ponderată *ajustată* (11). În aşa caz, nu se consideră valorile absolute ale estimărilor, suma ponderată ia în considerare devierile acestor valori de la estimarea medie a consumatorului respectiv. Altă metodă de a evita diferenţele în utilizarea scării de evaluare constă în filtrarea, bazată pe preferinţe: se supun pronosticului preferinţele utilizatorilor şi nu valorile absolute ale estimărilor. În SR colaborative se utilizează diferite metode de calcul al similarităţii *sim(c,c')* a consumatorilor. În majoritatea acestor metode similaritatea a doi consumatori se bazează pe faptul ce evaluări au făcut ei unora şi aceloraşi bunuri.

Metoda corelaţională şi metoda asemănării liniare au o răspândire largă. Fie $S_{x,y}$ – mulţimea bunurilor estimate atât de consumatorul **x**, cât şi de consumatorul **y**:

$$S_{x,y}=\{s \in S \lfloor r_{x,s} \neq \emptyset \ \& \ r_{y,s} \neq \emptyset\}.$$

În SR colaborative, S_{xy} se utilizează ca un rezultat intermediar la calculul aşa-numiţilor „vecini apropiaţi" ai consumatorului **x** şi frecvent este determinat prin calculul direct al intersecţiilor S_x şi S_y. Metoda corelaţională pentru calculul similarităţii foloseşte coeficientul de corelaţie Pearson:

$$sim(x,y) = \frac{\sum_{s \in S_{s,y}} (r_{x,s} - \bar{r}_x)(r_{y,s} - \bar{r}_y)}{\sqrt{\sum_{s \in S_{x,y}} (r_{x,s} - \bar{r}_x)^2 \sum_{s \in S_{x,y}} (r_{y,s} - \bar{r}_y)^2}} \tag{12}$$

În metoda asemănării liniare, ambii consumatori **x** şi **y** pot fi reprezentaţi ca vectori ai spaţiului m-dimensional, unde m=⌊$S_{x,y}$⌊. Atunci similiritatea a doi vectori poate fi determinată ca cosinusul unghiului dintre ei:

$$sim(x,y)=\frac{\vec{x}\cdot\vec{y}}{|\vec{x}|\cdot|\vec{y}|}=\frac{\sum_{s\in S_{x,y}} r_{x,s}\cdot r_{y,s}}{\sqrt{\sum_{s\in S_{x,y}} r_{x,s}^2}\sqrt{\sum_{s\in S_{x,y}} r_{y,s}^2}}, \quad (13)$$

unde $\vec{x}\vec{y}$ – produsul scalar a doi vectori.

Atragem atenţia că diferite SR pot folosi diferite abordări pentru un mod cât mai efectiv de determinare a similitudinii consumatorilor şi analizei estimărilor obţinute. Conceptul general acceptat constă într-un calcul preliminar al corespondenţelor între toţi utilizatorii sistemului şi revizuirea lor periodică (aşa necesitate nu apare frecvent, deoarece comunitatea utilizatorilor „vecini" este destul de constantă şi nu se schimbă radical în perioade mici de timp).

Menţionăm că atât metodele colaborative, cât şi cele de content folosesc aceeaşi măsură a similitudinii liniare (cosinusul unghiului dintre vectori), fapt recomandat de literatura de căutare a informaţiei. Cu toate acestea, SR de content utilizează măsura similitudinii liniare între vectorii ponderii frecvenţei/frecvenţei inverse, iar în SR colaborative similitudinea liniară se măsoară între vectorii estimărilor efectuate, specifice consumatorului.

Abordarea probabilistă a filtraţiei colaborative. În [2] se propune o abordare probabilistă a filtraţiei colaborative, unde estimările necunoscute se calculează astfel:

$$r_{c,s}=E(r_{c,s})=\sum_{i=0}^{n} i\cdot P(r_{c,s}=i\,|\,r_{c,s'}, s'\in S_c), \quad (14)$$

unde expresia probabilistică înseamnă că consumatorul **c** va da o notă sau alta bunului **s**, reieşind din estimările precedente, iar evaluările sunt numere întregi, de la 0 la n. Pentru a aprecia această probabilitate, în [2] se propun două modele probabilistice alternative: *modele claster* şi *reţele Bayes*.

În primul model consumatorii sunt grupaţi în clastere. Reieşind din apartenenţa consumatorului unui sau altui claster, evaluările lui sunt considerate independente, adică se poate aplica modelul clasic Bayes. Numărul de clastere şi parametrii modelului sunt determinaţi de baza de date.

Al doilea model reprezintă fiecare bun ca unul din nodurile reţelei Bayes, în care poziţia fiecărui nod corespunde estimării presupuse a bunului. Structura reţelei şi probabilităţile unor sau altor rezultate depind de baza de date.

În literatura de specialitate întâlnim şi alte abordări. Modelul statistic al filtrării colaborative a fost analizat în [10], unde au fost comparaţi diferiţi algoritmi de evaluare a parametrilor modelului (clasterizarea k-dimensională şi selecţia Gibbs). Shani [9] aplică modelul soluţiilor succesive pentru a elabora recomandări şi propune aplicarea lanţurilor Markov în acelaşi scop. Kumar [6] utilizează un model probabilistic simplu, pentru a demonstra că filtraţia colaborativă lucrează şi în cazul unei cantităţi relativ mici de informaţie despre consumator şi că în unele situaţii, în lipsă de informaţie suficientă, algoritmii simpli de filtrare colaborativă sunt la fel de efectivi, ca şi cei mai buni algoritmi bazaţi pe analiza utilităţii. În [12] se propune folosirea abordărilor probabilistice în calitate de metodă de reuniune a metodelor de model şi a celor anamnestice. În particular, se propune

- abordarea studiului activ ce formează modele probabilistice pentru preferinţele fiecărui consumator;
- în modelele complexe să se refere la profilurile păstrate ale consumatorilor, pentru a pronostica recomandări.

Metoda propusă dezvoltă ideile algoritmilor tradiţionali anamnestici. Alte modele probabilistice de filtrare colaborativă includ analiza Bayes, modelul relativist de probabilitate, modelul regresiei liniare, modelul entropiei maxime.

Unele SR folosesc metode hibride ce combină abordările colaborative şi de content, ceea ce permite de a evita restricţiile specifice fiecărui SR aparte. Reieşind din modul de a combina mecanismele de content şi colaborativitate, metodele hibrid pot fi grupate în următoarele clase:

- realizarea aparte a algoritmilor de content şi colaborativi şi reuniunea pronosticurilor lor;
- încorporarea unor reguli de content în metodica colaborativă;

- încorporarea unor reguli colaborative în metodica de content;
- elaborarea unui model comun ce cuprinde regulile ambelor metodici.

Abordarea [1] sugerează implementarea analizei Bayes cu aplicarea lanţurilor Markov şi a metodei Monte-Carlo. În procesul analizei, conform metodei Monte-Carlo calculatorul foloseşte o procedură de generare a numerelor aleatoare pentru imitarea datelor selecţiei generale studiate. Modelarea prin ecuaţii structurale alege eşantioane din selecţia generală conform indicaţiilor consumatorului, apoi efectuează următoarele acţiuni:

- simulează un eşantion aleatoriu din selecţia generală;
- efectuează analiza rezultatelor;
- salvează rezultatele.

După un număr mare de iteraţii, rezultatele păstrate reflectă bine repartiţia reală a statisticii de selecţie.

Metoda Monte-Carlo permite obţinerea informaţiei şi în cazurile când teoria repartiţiilor selective este inutilă. În particular, în [1] informaţia referitor la profilurile consumatorilor se foloseşte într-un model statistic unic ce studiază estimările necunoscute $\mathbf{r_{ij}}$ pentru consumatorul **i** şi bunul **j**.

$$r_{ij} = x_{ij}\,\mu + z_i\,\gamma_j + \omega_j\,\lambda_i + e_{ij}, \tag{15}$$

unde: $e_{ij} \sim N(0,\sigma^2)$, $\lambda_i \sim N(0, \Lambda)$, $\gamma_j \sim N(0,\Gamma)$ – variabile aleatoare ce iau în considerare erorile, nerecunoscute de izvoarele heterogenice cu repartiţii normale indicate; $i = \overline{1,I}$, $j = \overline{1,I}$ reprezintă, respectiv, consumatorii şi bunurile; x_{ij} – matricea ce conţine informaţia despre utilizatori şi bunuri; z_i – vectorul caracteristicilor consumatorului; ω_j – vectorul caracteristicilor bunurilor. Parametrii necunoscuţi ai acestui model – μ, σ^2, Λ şi Γ se determină după estimările deja primite utilizând lanţurile Markov şi metoda Monte-Carlo. Astfel, în [1] se folosesc parametrii consumatorului $\{z_i\}$ ce formează profilul lui, parametrii bunului $\{\omega_j\}$ ce formează profilul bunului şi interacţiunea lor $\{x_{ij}\}$ pentru analiza estimării bunului.

În plus, pe lângă metodele tradiţionale de elaborare a profilului de consum (cum ar fi bazarea pe cuvintele-cheie şi informaţia demografică), în ultimul timp au apărut metode noi ce au la bază prelucrarea automată a textelor (data-mining), analiza comportamentului de reţea etc.

Aceste metode permit să se ia în considerare interesele şi preferinţele consumatorilor şi, ca rezultat, să extindă profilul consumatorului. Asemenea metode se aplică şi pentru a extinde profilurile tradiţionale ale descrierii bunurilor, care până nu demult se limitau la descrierea cu ajutorul cuvintelor-cheie.

Odată ce profilurile bunului şi ale consumatorului au fost descrise, în caz general funcţia de prezicere a estimărilor poate fi definită prin aceste profiluri şi estimările anterioare în modul următor. Fie profilul consumatorului **i** se defineşte ca un vector de p atribute, $\vec{c} = (a_{i1}, a_{i2}, ..., a_{ip})$, profilul bunului **j** la fel se defineşte ca un vector de p atribute, $\vec{s}_j = (b_{j1}, b_{j2}, ..., b_{jr})$. În mod deliberat nu definim valorile atributelor $\mathbf{a_{ij}}$, $\mathbf{b_{ki}}$, deoarece acestea pot fi diferite în diferite aplicaţii (de exemplu, numere, categorii, reguli, traiectorii etc.). Fie $\vec{c}$ – vectorul tuturor consumatorilor, $\vec{c} = (\vec{c}_1, \vec{c}_2, ..., \vec{c}_m)$, iar $\vec{s}$ – vectorul tuturor bunurilor, $\vec{s} = (\vec{s}_1, \vec{s}_2, ...\vec{s}_n)$. Atunci, în general, procedura de prezicere a estimării se reprezintă prin

$$r'_{ij} = \begin{cases} r_{ij}, & dacă \quad r_{ij} \neq \varnothing \\ u_{ij}(R, \vec{c}, \vec{s}), & dacă \quad r_{ij} = \varnothing \end{cases} \tag{16}$$

Această funcţie prezice valoarea estimării r'_{ij} şi o exprimă prin estimări deja cunoscute şi profilurile $\vec{c}$ şi $\vec{s}$. Pentru a determina funcţia de utilitate u_{ij}, putem folosi diverse metode: euristice, clasificarea după cel mai apropiat „vecin", regresii, reţele neuronale, aproximări pe porţiuni, metode relaţionale etc.

Majoritatea SR colaborative folosesc algoritmul probabilistic Bayes sau algoritmul SVD (or SVD++). Dar, ambii algoritmi necesită un eşantion voluminos pentru studiu. În [3] a fost efectuată o analiză comparativă şi propus un algoritm alternativ ce nu cere eşantioane voluminoase. Compararea a fost făcută după criteriile de corectitudine a recomandărilor şi operativitatea.

STUDIA UNIVERSITATIS MOLDAVIAE, 2015, nr.2(82)
Seria "Ştiinţe exacte şi economice" ISSN 1857-2073 ISSN online 2345-1033 p.130-139

Algoritmul Bayes. Teorema Bayes [13] – una din teoremele de bază ale probabilităţii clasice, defineşte probabilitatea unui eveniment în condiţiile unei informaţii incomplete. Probabilitatea condiţionată a evenimentului A în condiţia realizării evenimentului B se notează P(A/B), conform [13]:

$$P(A/B)=\frac{P(A\cap B)}{P(B)},$$

unde P(A∩B) – probabilitatea realizării simultane a evenimentului A şi B, P(B) (P(A)) – probabilitatea realizării evenimentului B(A). Uşor de observat, că

$$P(A\cap B)=P(A|B)P(B)=P(B|A)P(A) \tag{17}$$

Din (17) rezultă

$$P(B|A)=\frac{P(A\cap B)}{P(A)} \tag{18}$$

Vom considera B – datele iniţiale (informaţia cunoscută), A – parametrii modelului supus studiului. De exemplu, B – datele referitor la ratingul bunurilor, determinat de consumatori, A – factorii supuşi studiului pentru consumatori şi bunuri. Fiecare probabilitate are sensul său: $P(B|A)$ – repartiţia probabilităţilor parametrilor modelului, după ce au fost luate în considerare datele iniţiale (probabilităţi aposteriori). $P(A|B)$ – aşa-numita verosimilitate. Neajunsul principal al algoritmului Bayes este cerinţa unor eşantioane voluminoase pentru analiză, şi, în situaţia când probabilitatea condiţională iniţială este egală cu zero, probabilitatea pronosticată la fel va fi egală cu zero. Din aceste motive, la moment sunt preferabili alţi algoritmi.

Algoritmul SVD. Se consideră o matrice R, ale cărei elemente sunt rating-uri (like-uri, cumpărături realizate etc.), pe care consumatorii (liniile matricei) au alocat bunurilor (coloanele matricei). De regulă, un consumator nu va estima un număr semnificativ de bunuri, de aceea matricea R va fi rarefiată puternic. Pentru aşa matrice de regulă se foloseşte descompunerea singulară [3] în forma $R=UDV^T$. R este o matrice de dimensiuni mari MxN, dar cu rangul mic şi poate fi descompusă în produs de matrice Nxf şi fxM, astfel micşorând numărul parametrilor de la NxM la (N+M)f.

Proprietatea de bază a SVD constă în aproximarea optimă, dacă în matricea D de lăsat primele f elemente diagonale, iar cele rămase de egalat cu zero.

$$X'UDV^T=\begin{pmatrix}\sigma_1\,0\ldots0\\0\;\;\sigma_2\ldots0\\\ldots\ldots\\0\;\;0\ldots\sigma_k\end{pmatrix}V^T\approx U\begin{pmatrix}\sigma_1\ldots0\;\;0\ldots0\\\ldots\ldots\ldots\\0\ldots\sigma_f\,0\ldots0\\0\ldots\;0\;\;0\ldots0\\\ldots\ldots\ldots\\0\ldots\;0\;\;0\ldots0\end{pmatrix}V^T \tag{19}$$

În matricea diagonală D, ce se află în mijlocul descompunerii singulare, elementele sunt ordonate: $\sigma_1\geq\sigma_2\geq\ldots\geq\sigma_k$, aşa că egalarea cu zero a ultimelor elemente înseamnă a anula cele mai mici elemente, iar f se alege reieşind din valorile singulare ale matricei. În cazul SR se obţine că fiecare consumator se reprezintă ca un vector din f factori U_i, iar fiecare bun – un vector din f factori V_j; apoi, pentru a prognoza rating-ul atribuit de consumatorul *i* bunului *j*, considerăm produsul lor scalar $U_iV_j=U_iV_j$. Problema se formulează astfel: având estimările cunoscute ale bunurilor, să prognozăm cât de bine va estima un consumator nou fiecare bun. Se introduc aşa-numiţii predictori de bază $b_{i,a}$ care se obţin din predictorii de bază ai fiecărui consumator aparte b_i, a unor bunuri aparte b_a, şi din rating-ul mediu general a bazei μ: $b_{i,a}=\mu+b_i+b_a$, unde μ – rating-ul mediu general al bazei, b_j – rating-ul mediu al consumatorului **i**, b_i – rating-ul mediu al consumatorului fiecărui bun **a**. Se cere de găsit aşa μ, b_i, b_a pentru predictorii de bază, pentru care $b_{i,a}$ cel mai

bine aproximează rating-urile date, apoi pot fi adăugaţi factorii. Dat fiind faptul că predictorii de bază au fost modificaţi, resturile sunt comparabile între ele, factorii se determină ca:

$$r^*_{i,a} = \mu + b_i + b_a + v_a^T u_i, \tag{20}$$

unde v_a – vectorul factorilor ce reprezintă bunul **a**, u_i – vectorul factorilor ce reprezintă consumatorul **i**. Acum problema poate fi formulată mai exact: trebuie de găsit cei mai buni predictori, care aproximează $r^*_{i,a}$, adică de minimizat eroarea:

$$L(\mu, b_i, b_a, v_a, u_i) = \sum_{(i,a)\in D} (r_{i,a} - r^*_{i,a})^2 = \sum_{(i,a)\in D} (r_{i,a} - \mu - b_a - v_a^T u)^2. \tag{21}$$

Funcţia L se minimizează prin metoda gradientului.

Algoritmul RNSA (The Refined Neighbor Selection Algorithm). Cu ajutorul algoritmului de clasterizare a K–mediilor se formează K clastere, fiecare constând din clienţi ce au preferinţe similare. Iniţial se alege un consumator arbitrar, iar în calitate de punct iniţial al centrului clasterului – k. Atunci fiecărui consumator i se atribuie un claster, astfel încât distanţa dintre el şi k să fie maximă. În calitate de distanţă poate fi utilizat coeficientul de corelaţie Pearson:

$$w_{u,a} = \frac{\sum_{i\in I} (r_{a,i} - \bar{r}_a)(r_{u,j} - \bar{r}_u)}{\sqrt{\sum_{i\in I} (r_{a,j} - \bar{r}_a)^2 \sum_{i\in I} (r_{u,j} - \bar{r}_y)^2}}, \tag{22}$$

unde $w_{u,a}$ – indică cât de aproape sunt consumatorii **u** şi **a**, I – mulţimea bunurilor estimate atât de **u**, cât şi de **a**; $r_{a,i}$, $r_{u,i}$ – evaluările bunului *i* de consumatorul **a**, respectiv **u**, $\bar{r}_a$, $\bar{r}_u$ – evaluările medii date de **a** şi **u**.

După finisarea clasterizării se alege clasterul cu cel mai înalt coeficient de corelaţie Pearson. Se calculează pronosticul pentru toţi consumatorii în acest claster. Etapele principale ale algoritmului RNSA sunt:

- Intrare: clientul-test t, datele iniţiale S
- Ieşire: vecini

1. Crearea a K clastere din S prin metoda clasterizării.
2. De găsit cel mai bun claster C pentru t.
3. De adăugat t în C şi de considerat ca v.
4. De adăugat v în lista vecinilor.
5. Dacă lista vecinilor este suficientă, întoarcem lista vecinilor. În caz contrar, extragem v din C şi pornim un proces de căutare în lăţime. Asemănarea consumătorului se verifică conform (22), dacă ea primeşte valori într-un diapazon fixat, atunci v – consumator şi se trece la pasul 4.

La pasul 5 algoritmul se opreşte, dacă numărul consumatorilor în căutare în lăţime este mai mare decât un careva număr fixat. Acest număr se determină prin diferite experimente. Formula de calcul al pronosticului

$$P_{a,j} = A(\bar{r}_{a,i}) + \frac{\sum_k (w_{a,k}(r_{k,j} - A(\bar{r}_{k,j})))}{\sum_k |w_{a,k}|}, \tag{23}$$

unde $P_{a,j}$ – pronosticul estimării, $w_{a,k}$ – gradul de apropiere între utilizatorii *a* şi *k*, $A(\bar{r}_{a,j})$, $A(\bar{r}_{k,j})$ – estimările medii date de utilizatorii *a* şi *k* respectiv.

Algoritmul RNSA se recomandă în cazul unor date iniţiale insuficiente şi permite obţinerea rezultatelor mai exacte. În cazul eşantioanelor mari se recomandă utilizarea algoritmului SVD.

SR perspective bazate pe metodologii stocastice. Majoritatea metodelor de elaborare a SR poartă un caracter determinist, deşi natura oricărui SR în mod obiectiv este stocastică. În adevăr, cerinţele consumatorilor parvin aleatoriu, orice prelucrare durează un timp aleatoriu. Astfel, este necesară dezvoltarea metodelor de elaborare a SR bazate pe teoria probabilităţilor, teoria proceselor aleatoare, teoria aşteptării, statisticii matematice etc. Nu se are în vedere aplicarea formulelor Bayes, care permit simplu de a recalcula probabilităţile a priori, după ce a avut loc un eveniment, ceea ce se foloseşte azi. La fel nu se iau în considerare SR

construite cu utilizarea lanţurilor Markov [13]. Indiscutabil, la o etapă anumită aceste metode şi-au îndeplinit rolul şi au contribuit esenţial la dezvoltarea SR, acum pot fi aplicate şi alte metode ale matematicii contemporane.

În SR pronosticul se elaborează în baza datelor referitor la consumator, care se obţin atât explicit, cât şi implicit. Această informaţie generează o multitudine de date noi, pe care SR contemporane le ignorează, deşi informaţia suplimentară, fiind prelucrată şi formatizată, poate schimba cardinal SR şi să aducă la elaborarea a noi SR, mai flexibile, mai adecvate proceselor reale. Această afirmaţie poate fi ilustrată [4] în baza exemplului de formalizare a fluxului cererilor consumatorilor.

Fie t_o, $t_1,\ldots t_n$, momente de timp (aleatoare) în care se înregistrează cererile consumatorilor, $z_k = t_k - t_{k-1}$, $k=1,2,\ldots,n,\ldots,-z_k$ denotă intervalul de timp între două cereri consecutive şi reprezintă o variabilă aleatoare continuă, $F_k(t)$ – funcţia de repartiţie a variabilei aleatoare z_k, $F_k(t)=P(z_k < t)$. Astfel, setul de funcţii $F_k(t)$ formează fluxul cererilor consumatorilor. Se consideră că fluxul de cereri este definit, dacă pentru orice *n* este definită repartiţia vectorului aleatoriu (z_1, z_2, …, z_n). Cu alte cuvinte, fluxul este definit, dacă sunt date repartiţiile $F_i(t),\ i=\overline{1,n}$.

Pentru a reflecta această informaţie în SR, este necesară informaţia referitor la natura concretă a fluxului: dacă variabilele aleatoare z_k sunt independente în totalitate şi identic repartizate, atunci $F_1(t)=F_2(t)=\ldots=F_n(t)=F(t)$ (în aşa caz, fluxul se numeşte flux recurent şi este suficient de bine studiat). Foarte comod în utilizare este cazul $F(t)=1-e^{-\lambda t}$, unde λ – constantă, λ>0 – aşa-numitul flux Poisson. Pentru fluxurile Poisson se poate a priori prognoza numărul de cereri pe o perioadă fixă de timp. Conform [4], probabilităţile $P_n(t)$ în intervalul de timp [0, t] sunt egale cu

$$P_n(t)=\frac{(\lambda t)^n}{n!}e^{-\lambda t}.$$

Menţionăm că aceste probabilităţi se calculează simplu şi că şi alte proprietăţi ale fluxului Poisson pot fi considerate în viitoarele SR.

În SR poate fi considerată şi dependenţa de trafic. Fie B – timpul de prelucrare a unei comenzi, ρ – coeficientul traficului, atunci

$$\rho=\frac{M(B)}{M(Z_k)},$$

unde $M(B)=\int_0^\infty tdB(t)$, $B(t) = P(B < t)$, $M(z_k)=\int_0^\infty td(A_k(t))$, $A_k(t) = P(z_k < t)$. Este cunoscut [4] că dacă $\rho < 1$, trafic nu există, iar pentru $\rho \geq 1$, SR vor lucra în condiţii de trafic intens. Acest exemplu arată cum pot fi considerate informaţii suplimentare în elaborarea a noi modele de SR mai flexibile şi mai adecvate proceselor reale.

Modele de SR cu pronostic virtual. În [15] a fost studiat un SR bazat pe aparatul transformării Fourier. Transformarea Fourier este bine cunoscută şi are aplicaţii în fizica matematică, mecanică etc. În opinia noastră, pentru elaborarea SR mai productive se vor utiliza transformarea Laplace şi Laplace-Stieltjes, care se determină astfel:

$$\overline{\alpha}(s)=\int_0^\infty e^{-sx}A(x)dx, \qquad \text{(I)}$$

$$\alpha(s)=\int_0^\infty e^{-sx}dA(x) \qquad \text{(II)}$$

În expresia (II) integrala se înţelege în sensul lui Stieltjes, iar funcţia A(t) – ca funcţie de repartiţie a unei variabile aleatoare. Una din proprietăţile principale ale acestor transformări este descrisă de teoremele Tauber. Fie $P(z,t)=\sum_k kP_k(t)$ – funcţia generatoare de repartiţi, $P_k(t)$ – numărul de comenzi în momentul de timp t, $\rho(z,t)=\int_0^\infty e^{-st}P(z,t)dt$ – transformata Laplace a funcţiei P(z,t). Vom nota cu P(z) funcţia

generatoare staţionară. Una din afirmaţiile teoremei Tauber constă în următoarele: în condiţiile unui regim staţionar (dacă coeficientul de încărcare ρ este mai mic ca unu)

$$P(z) = \lim_{s \to 0} \rho(z, s),$$

ultima egalitate stabileşte legătura dintre repartiţia numărului de cereri în orice moment de timp şi numărul de cereri în regim deja stabilit de funcţionarea sistemului. Această dependenţă va ajuta la fundamentarea şi va extinde domeniul de extrapolare a soluţiilor sistemelor de recomandare. În plus, prin integrala Stieltjes putem atribui un sens probabilistic transformatei Laplace-Stieltjes. În teoria sistemelor de deservire aceasta metodă este cunoscută sub denumirea „metoda catastrofelor" [14] şi poate fi utilizată pentru determinarea strategiei SR.

Modele de SR cu priorităţi incluse. Este normal ca în anumite condiţii unele cereri sau clase de cereri să posede anumite privilegii în deservire. Măsura acestor privilegii se numeşte prioritate. Prioritatea poate fi introdusă atât din exteriorul sistemului, cât şi din interiorul lui, în dependenţă de anumite circumstanţe. Mai mult, pot fi considerate priorităţi dinamice – priorităţi ce se schimbă în dependenţă de starea sistemului. Dezvoltarea SR cu priorităţi poate fi justificată şi de faptul că în [4] a fost demonstrat că din toate modelele SR legate de deservirea cererilor, deservirea cu priorităţi este optimală. O clasă largă de priorităţi – discipline generalizate de prioritate – a fost introdusă în [14]. Unele discipline noi de prioritate dinamică de clasă DD (Discretional Dynamical) au fost recent studiate în [5].

Concluzii

Avantajul disciplinelor de prioritate generalizate constă în faptul că modelele de deservire înzestrate cu asemenea priorităţi permit a considera diferite pierderi de timp necesare pentru a schimba modul de deservire, a restabili informaţia, a îndeplini unele operaţii auxiliare etc. ce au loc în sistemele reale.

Cel mai important avantaj al SR cu priorităţi constă în semimarkovitatea disciplinelor de prioritate generalizate, ceea ce înseamnă că ele sunt invariante în raport cu tipul funcţiilor de repartiţie. Aceasta le face extrem de atractive şi importante pentru diverse aplicaţii, deoarece ele permit modelarea unui spaţiu larg de probleme, inclusiv evoluţia diferitelor moduri de funcţionare a sistemelor reale.

Bibliografie:

1. ANSARI, A., ESSEGAIER, S. and KOHLI, R. „Internet Recommendations Systems". In: *J. Marketing Research*, 2000, p.363-375.
2. BREESE, J.S., HECKERMAN, D. and KADIE, C. „Empirical Analysis of Predictive Algorithms for Collaborative Filtering". In: *Proc. 14th Conf. Uncertainty in Artificial Intelligence*, 1998.
3. GEDIMINAS, A., TUZHILIN, A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions IEEE. In: *Transactions on Knowledge and Data Engineering*, 2005, vol.17, no.6.
4. MISHKOY, Gh. *Textbook on Queueing Analysis*. ULIM, 2008.
5. MISHKOY, Gh., MITEV, L. Performance Characteristics for DD Priority Discipline with Semi-Markov switching. In: *Communications in Computer and Information Science*. Springer, 2014, p.2204-2018.
6. KUMAR, R., RAGHAVAN, P., RAJAGOPALAN S., and TOMKINS, A. Recommendation Systems: A Probabilistic Analysis. In: *J. Computerand System Sciences*, 2001, vol.63, no.1, p.42-61.
7. PAZZANI, M. and BILLSUS, D. Learning and Revising User Profiles: The Identification of Interesting Web Sites. In: *Machine Learning*, 1997, vol.27, p.313-331.
8. RICCI, F. *Systems Handbook*. Springer, 2011.
9. SHANI, G., BRAFMAN, R. and HECKERMAN, D. An MDP-Based Recommender System. In: *Proc. 18th Conf. Uncertainty in Artificial Intelligence*, 2002.
10. UNGAR, L.H. and FOSTER, D.P. *Clustering Methods for Collaborative Filtering.* Proc. Recommender Systems, Papers from 1998 Workshop, Technical Report WS-98-08, 1998.
11. WHITTLE, P. *Networks: Optimization and Evolution*. Cambridge Univ. Press, 2007.
12. YU, K., SCHWAIGHOFER, A., TRESP, V., XU, X. and KRIEGEL, H.-P. Probabilistic Memory-Based Collaborative Filtering. In: *IEEE Trans. Knowledge and Data Eng.*, 2004, vol.16, no.1, p.56-69.
13. МИШКОЙ, Г.К. *Вероятность и Математическая Статистика. Курс лекций*. Кишинэу: Elan Poligraf, 2012.
14. МИШКОЙ, Г. *Обобщенные приоритетные системы.* Академия Наук Молдовы, 2009.
15. ШТЕПИКОВ, Д.Г. и др. *Алгоритмы рекомендаций, основанные на Фурье-анализе профилей пользователей*. Санкт-Петербург: Телематика, 2009.
16. http://www.machinelearning.ru/wiki/

Prezentat la 01.06.2015