Analiză bibliometrică a domeniului sistemelor de recomandare (1980–2026): structură tematică, rețele de colaborare și tendințe

Introducere

Contextul general al lucrării pornește de la poziția clară că sistemele de recomandare au devenit componente esențiale ale produselor digitale contemporane — de la platformele de e-commerce (de ex. Amazon), la servicii de streaming (Netflix, Spotify) și rețele sociale cu audiențe masive (Facebook, TikTok, YouTube) — contribuind decisiv la personalizarea ofertelor, reducerea supraîncărcării informaționale și creșterea retenției utilizatorilor. Literatura de specialitate reflectă această importanță printr-un palier larg de metode, de la filtre colaborative și content-based până la modele hibride, iar în ultimii cinci ani cercetarea a înregistrat o tranziție clară către tehnici avansate de reprezentare (contrastive şi self-supervised learning), arhitecturi pe grafuri (graph neural networks) și soluții distribuite (federated learning), tendințe care sunt aplicate tot mai frecvent pe seturi mari de date industriale. Deși studiile bibliometrice anterioare au contribuit cu hărți utile ale domeniului — adesea realizate cu instrumente precum VOSviewer sau CiteSpace pentru analiza co-citărilor și a co-ocurențelor — ele au fost în multe cazuri limitate ca interval temporal sau dimensiune a corpusului; având la dispoziție o bază extinsă (≈16.000 de articole din Web of Science), analiza bibliometrică orientată către ultimii cinci ani devine esențială pentru a surprinde nu doar direcțiile metodologice emergente, ci și reconfigurările tematice și structurale ale comunității științifice, informații critice pentru prioritizarea cercetării și pentru deciziile editoriale și de finanțare.

Ce s-a mai făcut anterior include atât progresul metodologic — de la filtre colaborative și sisteme bazate pe conținut la modele hibride și arhitecturi de deep learning —, cât și numeroase sinteze tematice focalizate pe subdomenii. În ultimii cinci ani, accentul cercetării s-a mutat vizibil către tehnici de reprezentare (de exemplu contrastive learning, self-supervised learning), rețele graf (graph neural networks) și soluții distribuite (federated learning), iar lucrările au alternat între dezvoltarea de metode noi și aplicarea practică în seturi mari de date.

Topicul ridică provocări tehnice și metodologice notabile: sparsitatea și cold-start-ul pentru utilizatori şi itemi, evaluarea robustă și reproductibilitatea experimentelelor, balansul între acuratețe și explicabilitate, precum și problema biasului și a confidențialității în scenarii personalizate. La acestea se adaugă provocări bibliometrice — fragmentarea terminologică și variația calității metadatelor — care îngreunează comparațiile la scară largă între studii.

Motivația studiului este pragmatică: vreau să fac o analiză bibliometrică ca să văd cum a evoluat topicul de-a lungul ultimilor 5 ani, pentru a identifica direcțiile emergente, ariile saturate și legăturile structurale din comunitatea de cercetare. O imagine bazată pe date oferă context obiectiv pentru prioritizarea eforturilor de cercetare și pentru formularea de recomandări privind practici de publicare și colaborare.

Obiectivele acestui research sunt clar definite: ne propunem să cartografiem evoluția tematică recentă, să identificăm subdomeniile cu creștere rapidă, să descriem structura colaborativă a autorilor și a instituțiilor, să analizăm distribuția impactului pe baza citărilor și să evaluăm tonul comunicării științifice prin analize de sentiment ale abstractelor. Aceste obiective se traduc în analize concrete — calculul CAGR pentru cuvinte-cheie, modelarea de topic, analiza rețelelor de co-autorat, diagrame Sankey pentru fluxurile afiliație→jurnal și evaluarea distribuției citărilor.

Pornim de la câteva ipoteze de lucru: ipoteza 1 presupune că conceptele asociate tehnicilor moderne de reprezentare au o creștere semnificativă în ultimii 5 ani; ipoteza 2 susține că impactul științific este puternic concentrat în jurul unui număr mic de lucrări; ipoteza 3 afirmă că rețelele de colaborare prezintă o modularitate ridicată, cu autori-punte care leagă comunități relativ izolate.

Există și controverse relevante pentru interpretare: unele sisteme au deficiențe sau neajunsuri — de la degradarea performanței în condiții reale la riscuri etice precum perpetuarea biasului — astfel încât pluralitatea arhitecturilor și a abordărilor reflectă în parte încercările de a remedia aceste limite. Din perspectivă bibliometrică, controversa se extinde la modul în care evaluăm „impactul": dependența exclusivă de citări poate ascunde calități practice esențiale sau poate favoriza anumite grupuri de cercetare.

Întrebările de cercetare pe care le urmărim sunt formulate astfel: care sunt temele emergente și cele aflate în regres în intervalul analizat; în ce măsură structura colaborativă influențează difuzarea ideilor; cum este distribuit impactul științific între lucrări și ce implicații are această concentrare pentru evaluarea cercetării; ce relații instituționale se pot observa între afiliații autorilor și jurnalele alese pentru publicare; și cum variază tonul comunicării științifice în timp, pe baza analizelor de sentiment ale abstractelor.

Contribuția și noutatea lucrării constau în combinarea analitică a metodelor bibliometrice și NLP aplicate pe un corpus extins și actualizat: raportăm rezultate mai importante, mai originale prin calculul sistematic al CAGR pe top-100 de cuvinte-cheie, identificarea clusterelor tematice (cu atenție critică la redundanțele terminologice), aplicarea LDA pe abstracte și cartografierea fluxurilor afiliație→jurnal via Sankey, oferind totodată recomandări practice pentru normalizarea termenilor și pentru îmbunătățirea designului studiilor viitoare.

Structura lucrării urmează un fir logic: în prima secțiune am făcut introducerea, în a doua secțiune prezentăm datele și metodologia — incluzând pașii de preprocesare și parametrii de modelare —, în a treia secțiune expunem rezultatele (statistici descriptive, analiza citărilor, rețele de co-autorat, clusterizare de cuvinte-cheie, LDA, heatmap, Sankey, analiza sentimentului și wordcloud), iar secțiunea finală oferă discuții critice, concluzii și recomandări metodologice și practice.

Rezultate

Analiza pornește de la un set extins de date extras din Web of Science și monitorizat pe parcursul lunii august 2025. Prima colectare a fost realizată la 1 august, iar snapshot-ul final, utilizat în toate analizele, a fost înregistrat la 30–31 august. Setul reunit în fișierul *merged_output.xlsx* cuprinde 15.944 de înregistrări, acoperă perioada 1980–2026 și include 72 de variabile, printre care titluri, autori, abstracte, cuvinte-cheie, surse, ani de publicare și număr de citări. În total, corpusul însumează 15.183 de autori unici, 6.580 de surse științifice și 176.910 citări.

Scopul este extragerea de informații relevante din această masă de date eterogenă, în condițiile în care doar informațiile procesate și interpretate pot sprijini concluzii valide. Variabilele au fost tratate diferențiat: cele numerice (ex. *Times Cited*, *Publication Year*) au permis calcule statistice și analize temporale, iar cele textuale (autori, titluri, cuvinte-cheie, abstracte, afilieri) au stat la baza analizelor conceptuale și relaționale.

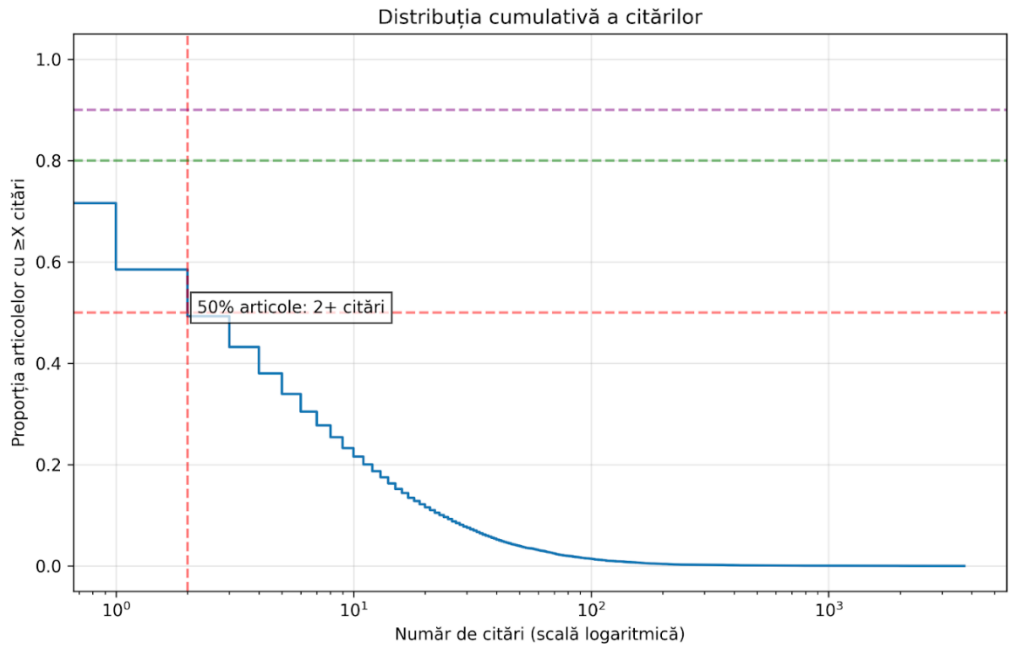
Preprocesarea a inclus verificarea valorilor lipsă, conversii de tip, normalizarea textuală, deduplicarea autorilor și unificarea variantelor lexicale pentru cuvinte-cheie. Rețeaua de co-autori a fost construită pe baza co-aparițiilor, evaluată prin degree și betweenness centrality, iar structura colaborativă a fost analizată prin algoritmi de detectare a comunităților. Pentru dinamica terminologică am aplicat calculul ratei de creștere anuală compusă (CAGR) pe top 100 de cuvinte-cheie și un heatmap temporal. Abstractele au fost procesate cu LDA (cinci topice), iar analiza sentimentului a fost realizată cu modelul VADER, interpretat cu precauțiile necesare pentru limbaj academic. Vizualizări complementare, precum wordcloud pe titluri și diagrama Sankey afiliere→jurnal, completează analiza.

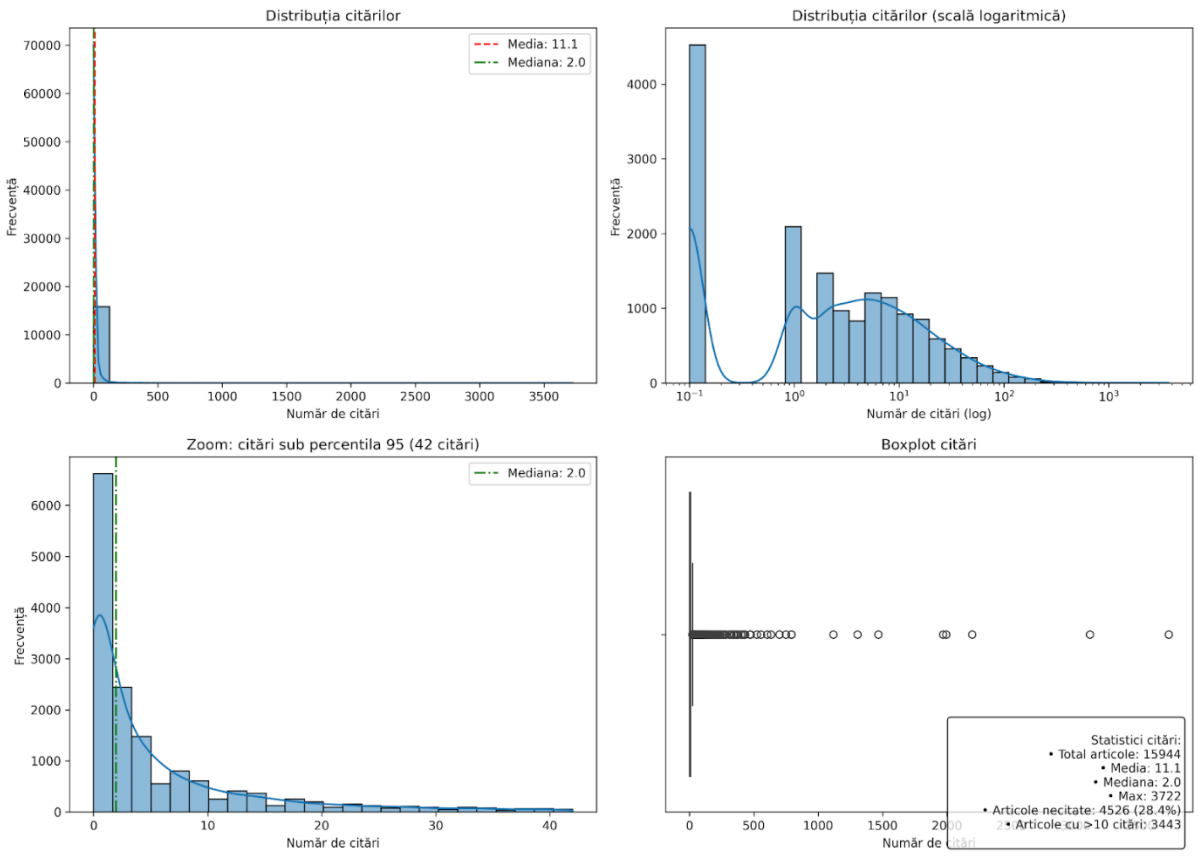
Statistica descriptivă confirmă un tipar caracteristic domeniilor emergente: media citărilor este 11,1, însă mediana este doar 2,0, ceea ce arată o concentrare a impactului în jurul unui număr mic de articole. Aproximativ 28,4% dintre publicații nu au primit nicio citare, în timp ce articolul cu cel mai mare succes a acumulat 3.722 de citări. Această asimetrie arată că, deși domeniul este prolific ca volum, influența științifică este puternic polarizată.

Pentru a asigura transparența și replicabilitatea, întregul set de date și scripturile aferente preprocesării și analizelor au fost puse la dispoziția comunității într-un repository public GitHub, menționat explicit în articol printr-o notă de subsol. Această deschidere permite verificarea calculelor și extinderea cercetării pe direcții noi.

Notă de subsol: Datele și codul asociat sunt disponibile public pe GitHub la adresa: https://github.com/ghineatudor/Analiza-bibliometrica-pe-sisteme-de-recomandare.git.

Distribuția citărilor

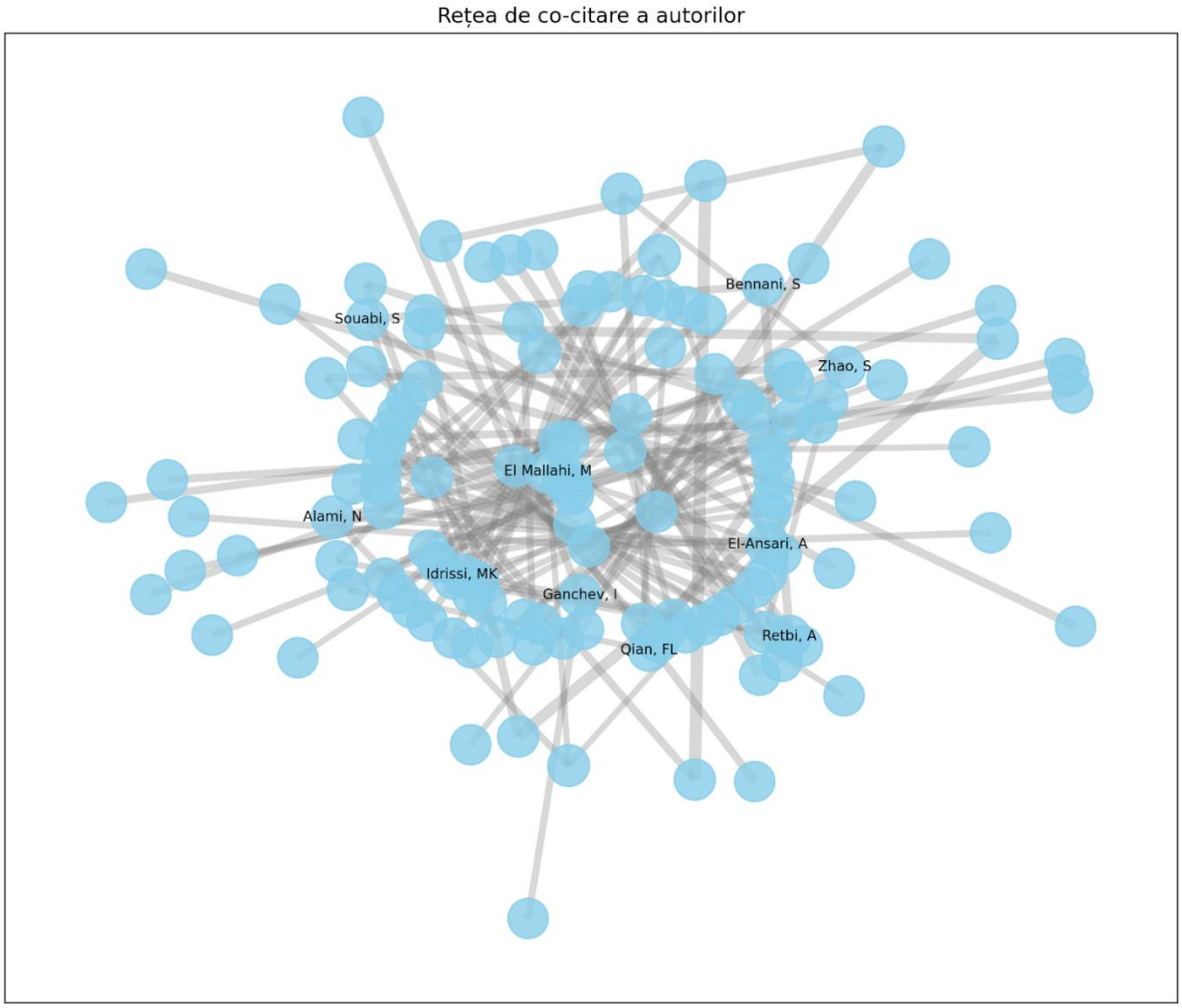




Analiza distribuției citărilor arată o asimetrie pronunțată spre dreapta: valorile extreme ale numărului de citări din vârf trag media în sus, în timp ce majoritatea articolelor rămân cu numere reduse de citări. Această formă a distribuției urmărește legea lui Pareto, conform căreia un procent mic de articole atrage majoritatea citărilor. În termeni practici, acest lucru

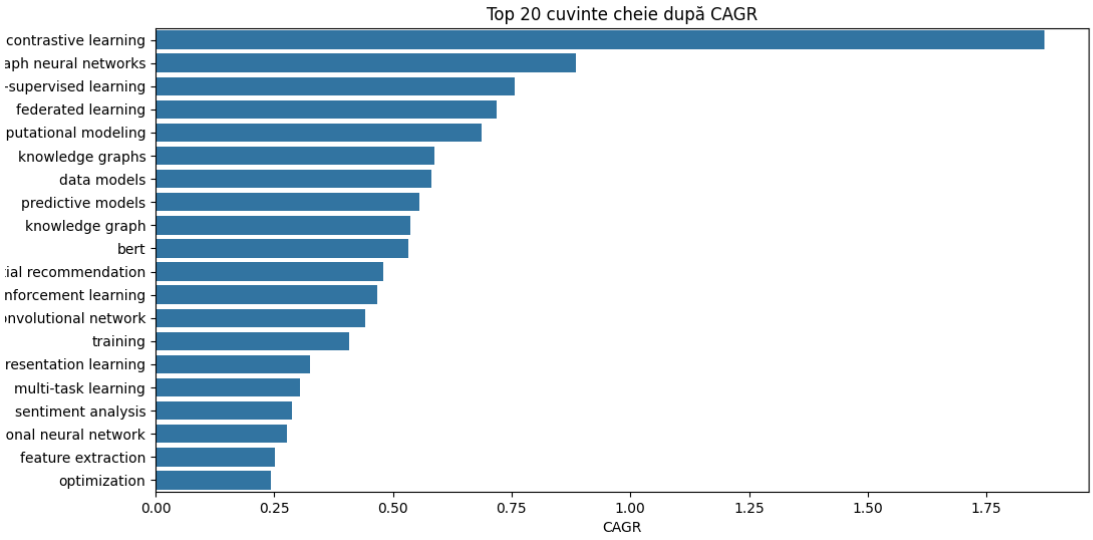
sugerează că evaluările bazate exclusiv pe medii sau pe numărul brut de publicații pot fi înșelătoare; este recomandabilă utilizarea de metrici robuste sau percentilare în evaluarea impactului științific.

Rețeaua de co-autori



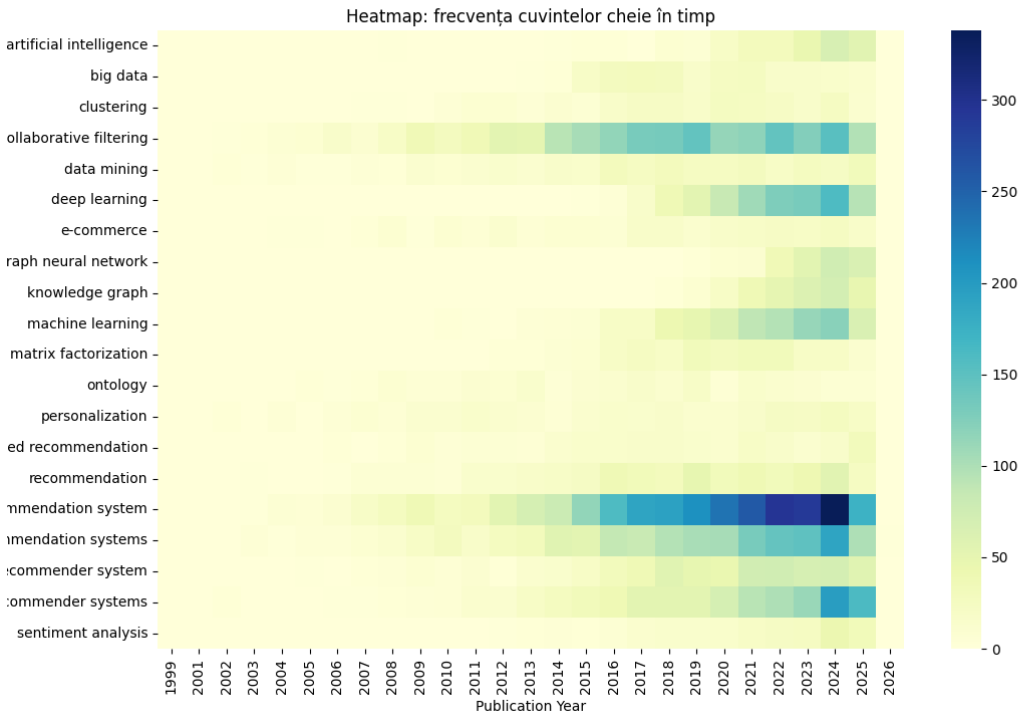
Rețeaua de co-autori dezvăluie existența a 53 de grupuri de colaborare cu o densitate generală scăzută, de 0,0112, ceea ce indică o conectivitate redusă la scară globală și apariția unor clustere bine definite, relativ izolate. Autorii cu centralitate ridicată identificați în analiză includ Alami N., El-Ansari A., El Mallahi M., Ganchev I. și Qian F.L., aceștia jucând roluri de punte între comunitățile de cercetare și facilitând transferul de idei. Structura rețelei sugerează existența unor „școli” de gândire foarte legate intern, dar relativ separate una de cealaltă, iar deschiderea către colaborări intercluster poate crește difuzarea noilor metode.

Creșteri tematice (CAGR)



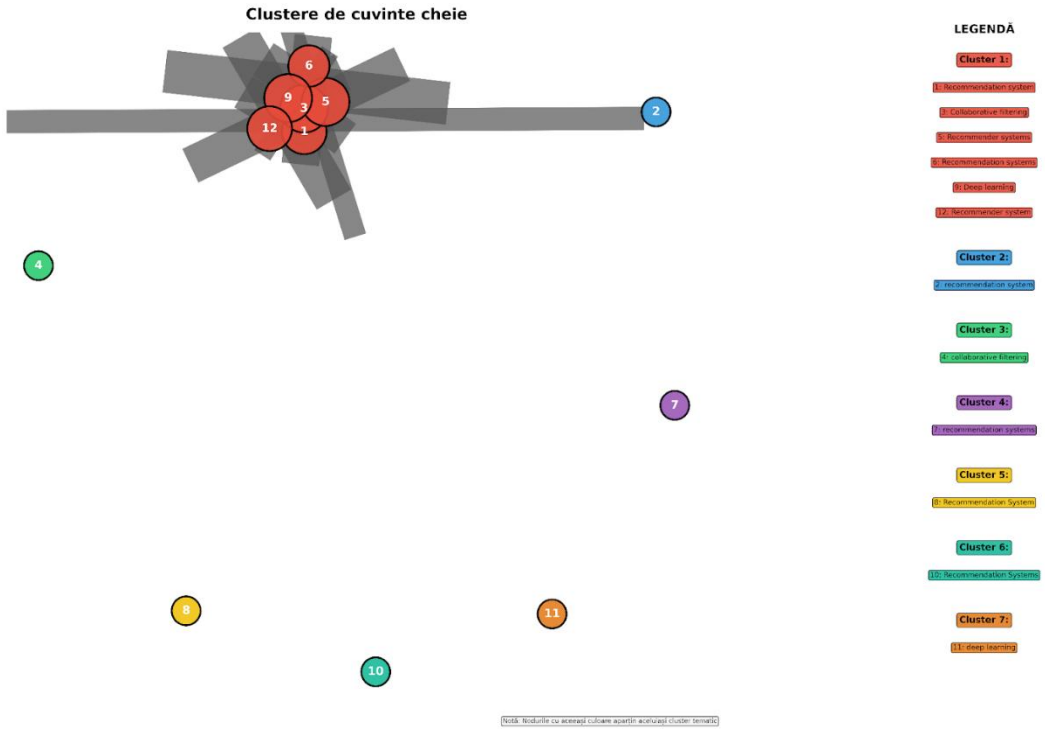
Calculul ratei de creștere anuală compuse pentru frecvența cuvintelor-cheie relevă direcții tematice în expansiune și altele în regres. Cel mai mare CAGR identificat aparține termenului contrastive learning, cu o valoare de aproximativ 1,8716, ceea ce semnalează o creștere rapidă a interesului pentru acest subdomeniu. Următoarele concepte cu CAGR semnificativ pozitiv sunt graph neural networks la 0,8860, self-supervised learning la 0,7556, federated learning la 0,7188 și computational modeling la 0,6864. În sens opus, termenii care înregistrează CAGR negativ includ large language models cu −0,4497, graph neural network (o intrare similară etichetată diferit) cu −0,1452, session-based recommendation cu −0,0943, k-means cu −0,0706 și transfer learning cu −0,0670. Aceste rezultate indică faptul că domeniul se îndreaptă puternic către tehnici moderne de învățare reprezentativă și sisteme distribuite, iar anumite metode sau denumiri își pierd din frecvență fie din cauza maturizării, fie din cauza rebranduirii conceptuale.

Heatmap temporal al cuvintelor-cheie



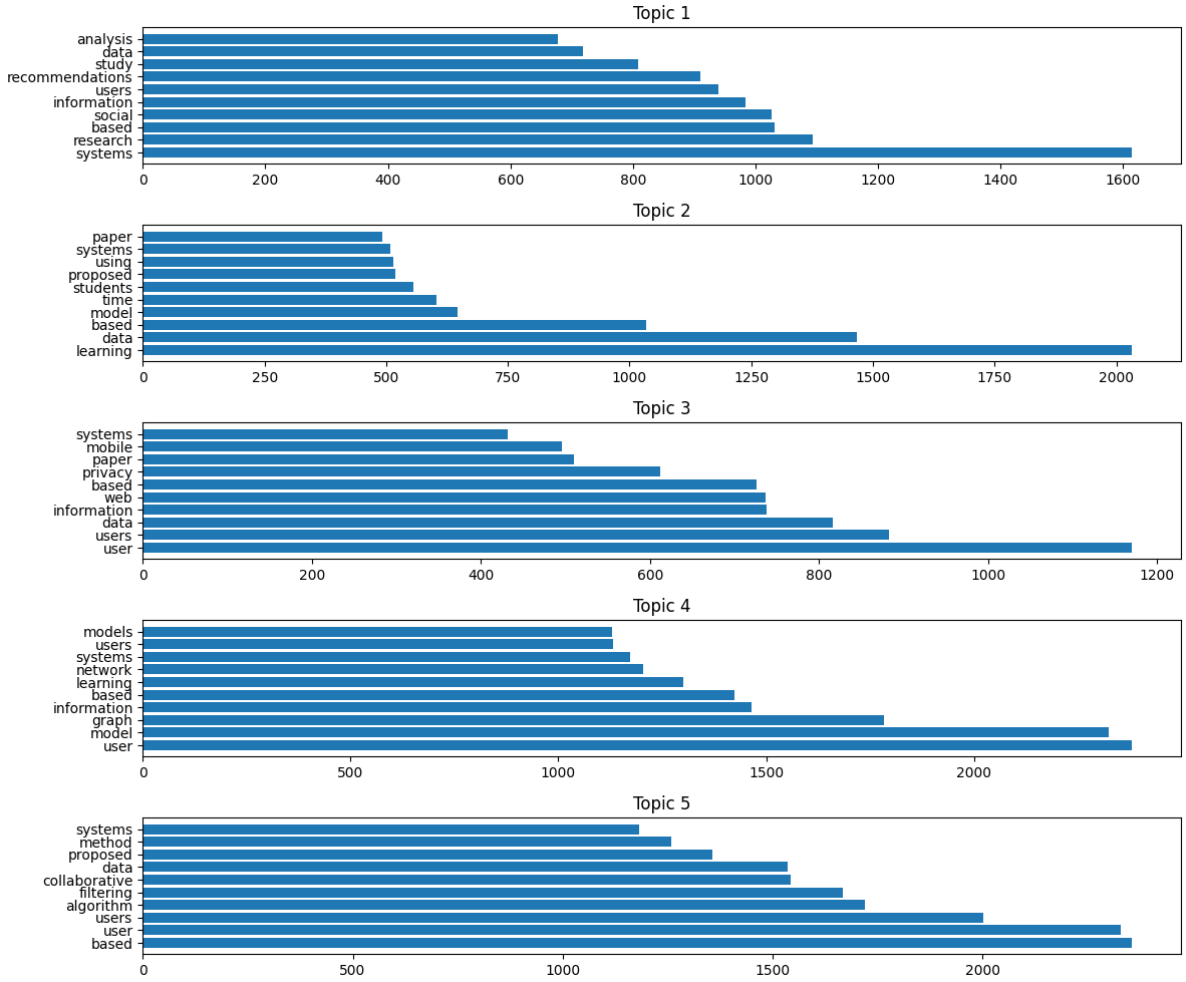
Heatmap-ul temporal ilustrează evoluția frecvenței unor termeni-cheie de-a lungul anilor. Zonele de intensitate mai mare corespund perioadelor în care anumite concepte au fost folosite frecvent. Termeni precum recommendation system, collaborative filtering și deep learning prezintă variații notabile de-a lungul timpului și domină perioade lungi, ceea ce indică că aceste concepte rămân nucleare în domeniu. Observația importantă este că, deși câteva concepte emergente au CAGR mare, ele nu se manifestă încă întotdeauna printr-o prezență stabilă la rândul întregului interval temporal; unele cresc rapid, dar într-un eșalon temporal mai restrâns.

Clusterele de cuvinte-cheie și componentele lor



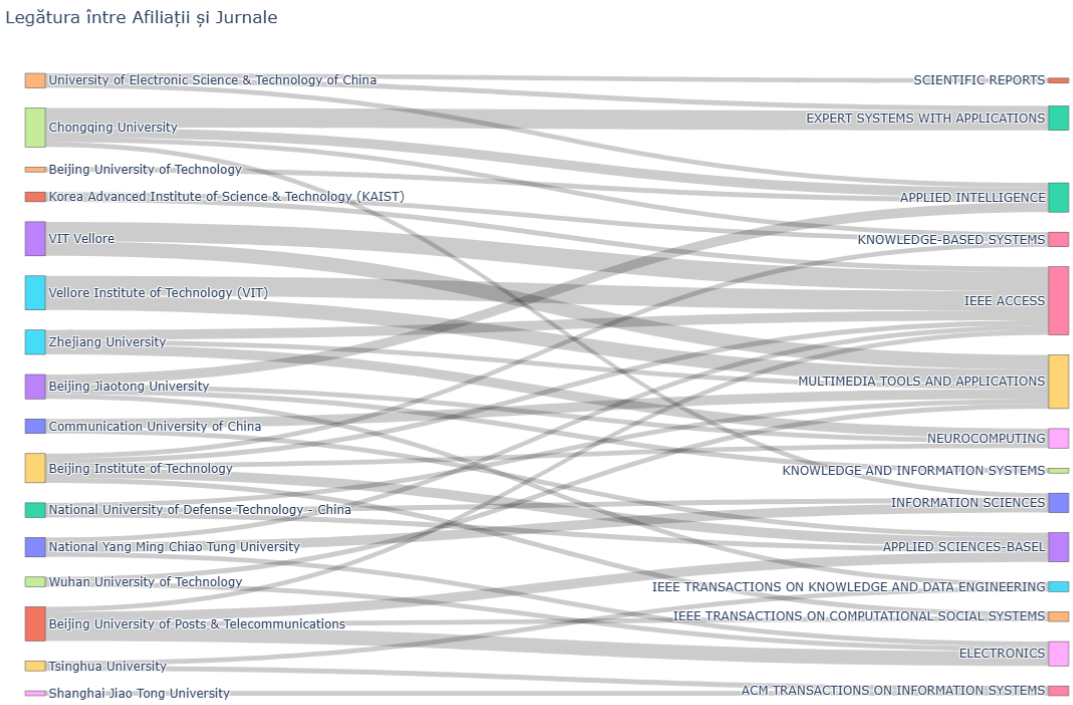
Analiza clusterelor de cuvinte-cheie a evidențiat un nucleu conceptual concentrat în jurul sistemelor de recomandare și al filtrării colaborative, dar a revelat și o redundanță lexicală semnificativă care fragmentează același concept în mai multe etichete. Astfel, componenta clusterelor poate fi redată succint: clusterul întâi include termeni precum Recommendation system, Recommender system, Recommendation systems, Deep learning și Collaborative filtering, clusterul doi este centrat pe termenul recommendation system, clusterul trei pe expresia collaborative filtering, clusterul patru pe recommendation systems, clusterul cinci pe forma Recommendation System, clusterul șase pe Recommendation Systems și clusterul șapte este dedicat deep learning. Această fragmentare denotă variații de capitalizare, pluralizare și sinonimie care ar trebui normalizate pentru o interpretare mai clară; în forma actuală, mai multe clustere reflectă practic aceeași tematică, ceea ce face dificilă distingerea subdomeniilor reale fără o etapă de deduplicare și normalizare lexicală.

LDA pe abstracte: topice și cuvinte reprezentative



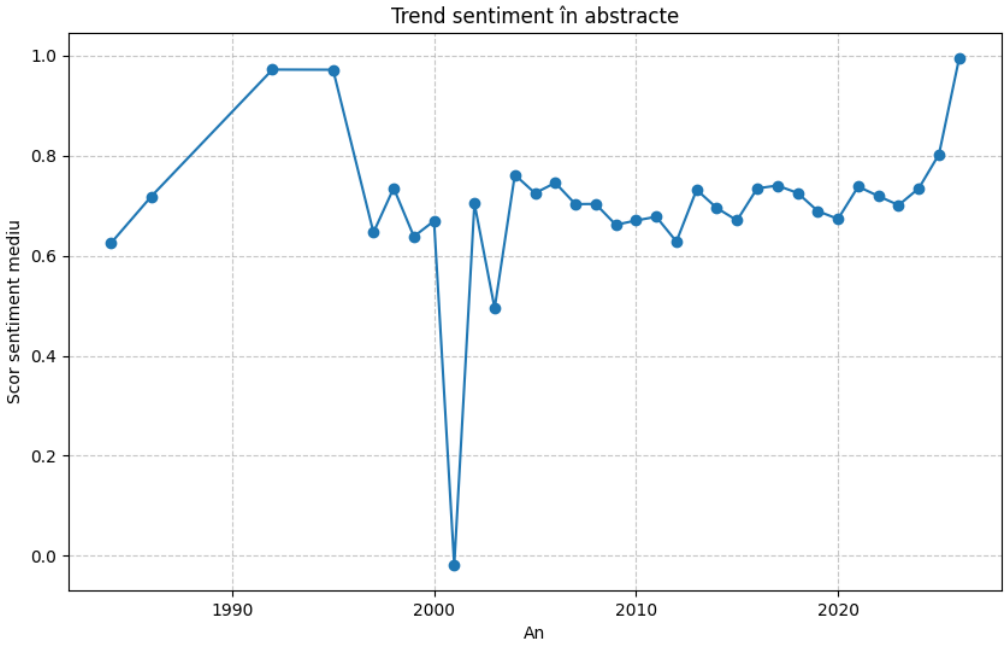
Aplicarea modelului LDA cu cinci topice pe setul de abstracte a generat distribuții de cuvinte care conturează principalele arii tematice ale corpusului. Primul topic este dominat de termeni precum analysis, data, study, recommendations, users, information, social, based, research și systems, sugerând un cluster orientat spre analize empirice și studii privind utilizatorii și datele. Al doilea topic cuprinde cuvinte ca paper, systems, using, proposed, students, time, model, based, data și learning, indicând lucrări metodologice sau aplicate, uneori pe teme educaționale sau de învățare în timp. Al treilea topic, în care apar tokens precum systems, mobile, paper, privacy, based, web, information, data, users și user, semnalează o preocupare pentru aplicațiile mobile, web și aspectele de confidențialitate. Al patrulea topic reunește termeni ca models, users, systems, network, learning, based, information, graph, model și user și reflectă interesul pentru modele de rețea, învățare și graph-based approaches. Al cincilea topic pune accentul pe systems, method, proposed, data, collaborative, filtering, algorithm, users, user și based, dezvăluind o tematică clasică de filtrare colaborativă și dezvoltare algoritmică. Observația metodologică este că anumite topice conțin termeni foarte generici precum systems sau based; o preprocesare textuală mai fină (stop-words specifice domeniului, lematizare, unificare de forme) ar spori claritatea și discriminativitatea topivelor.

Sankey: fluxul afiliație → jurnal



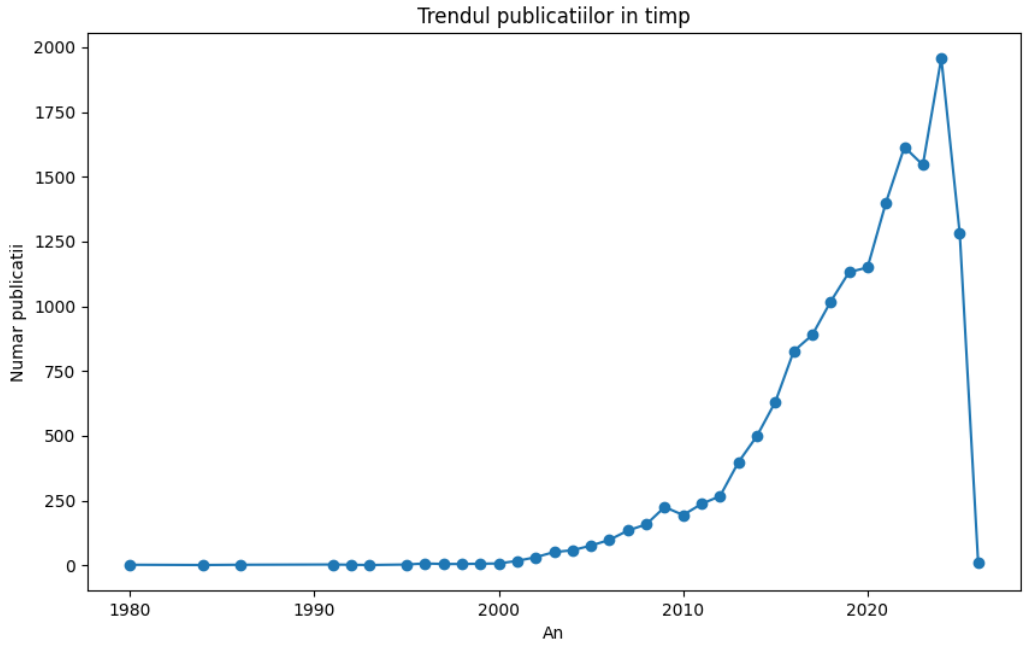
Analiza Sankey pentru fluxurile afiliație către jurnal relevă conexiuni notabile între anumite instituții și reviste. Printre legăturile principale identificate se numără Chongqing University către Expert Systems with Applications, cu 4 publicații în conexiunea respectivă, Vellore Institute of Technology (VIT Vellore) către IEEE Access cu 4 publicații și către Multimedia Tools and Applications cu 3 publicații, precum și Beijing University of Posts & Telecommunications către Electronics cu 3 publicații. Aceste canale preferențiale sugerează că anumite instituții aleg constant aceleași platforme pentru diseminarea rezultatelor lor, iar jurnalele menționate funcționează ca agregatoare tematice cu audiență largă în acest domeniu.

Analiza de sentiment pe abstracte



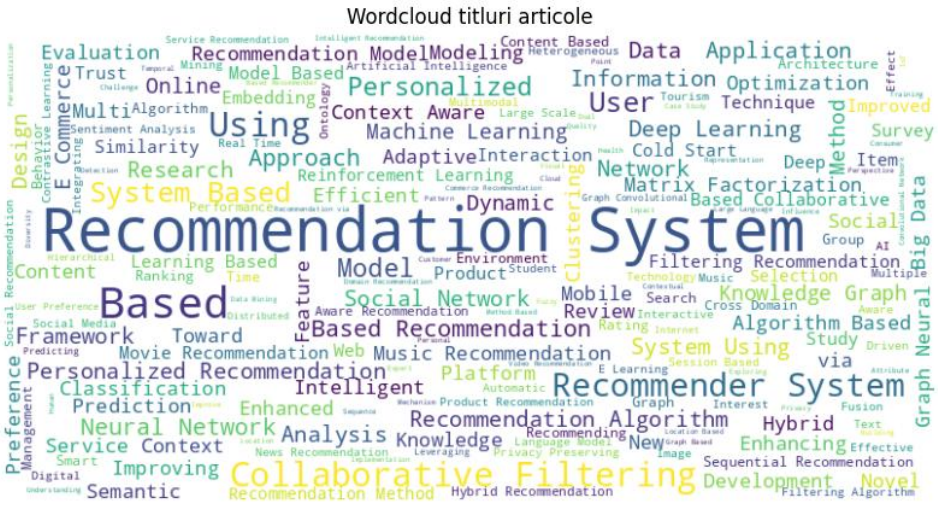
Agregarea scorurilor de sentiment pe ani, obținute prin instrumentul VADER, indică un sentiment mediu anual în jurul valorii de 0,718 pe scala −1..+1, ceea ce reflectă un ton general pozitiv în redactarea abstractelor din acest domeniu. Totuși, există variații anuale și valori atipice, cum ar fi 2001 cu un scor apropiat de zero negativ și 2026 cu un scor foarte ridicat, aproape de +1; aceste extreme pot fi cauzate de prezența unui număr mic de înregistrări pentru anii respectivi, erori sau limitările metodei VADER pe limbaj tehnic. Interpretarea acestor scoruri trebuie făcută cu prudență: un ton „pozitiv” în abstract nu înseamnă neapărat rezultate substanțial pozitive, ci mai degrabă un limbaj orientat către propuneri, contribuții și rezultate favorabile.

Trendul publicațiilor



Evoluția numărului anual de publicații arată o tendință ascendentă clară, cu vârfuri de activitate în anii 2022, 2023 și 2024. Comparativ cu anul inițial analizat, creșterea procentuală raportată în date este de aproximativ 300%, confirmând faptul că interesul pentru sistemele de recomandare a cunoscut o expansiune semnificativă, probabil corelată cu avansul metodelor de învățare automată și cu adoptarea largă a sistemelor de recomandare în industrie.

Wordcloud din titluri



Analiza frecvenței termenilor din titlurile articolelor arată predominanța cuvintelor recommendation (8.443 apariții) și system (4.197 apariții), urmate de based (3.172), learning (1.969), using (1.814), systems (1.461), collaborative (1.287), personalized (1.152), recommender (1.110) și filtering (1.103). Aceasta confirmă direcționarea majoră a cercetării către sisteme bazate pe învățare și filtrare colaborativă, iar termenii mai rar întâlniți pot indica nișe emergente sau direcții secundare de investigație.

Discuție

Rezultatele sintetizate indică un domeniu dinamic, cu nucleu conceptual stabil în jurul sistemelor de recomandare tradiționale și cu o tranziție pronunțată către tehnici moderne de învățare reprezentativă, inclusiv contrastive learning, metode self-supervised și graph neural networks. Fragmentarea terminologică descoperită în analiza clusterelor sugerează o nevoie urgentă de normalizare lingvistică înainte de efectuarea unor analize comparative serioase; fără această etapă, riscul este ca aceleași concepte să fie sparte artificial în mai multe grupuri. Concentrarea impactului în jurul unui număr mic de lucrări indică dificultăți pentru autori mai puțin conectați în a-și face cunoscute rezultatele, ceea ce relevă importanța strategiilor de vizibilitate și colaborare pentru creșterea impactului. În privința metodologiei, LDA clasic și VADER oferă insighturi utile, dar limitate; pentru rezultate mai robuste recomand utilizarea unor tehnici moderne de topic modeling pe embeddinguri și a unor modele de sentiment calibrate pentru limbaj academic.

Concluzii și recomandări practice

Analiza bibliometrică arată că sistemele de recomandare reprezintă un domeniu în creștere, dominat în continuare de teme tradiționale precum recommendation system și collaborative filtering, dar deschis unei schimbări metodologice consistente spre tehnici avansate de învățare automată. Pe baza observațiilor, recomand efectuarea normalizării termenilor (lowercase, lematizare, unificare sinonimică) înainte de refacerea clusterizării pentru a elimina

redundanțele evidente. Totodată, pentru a obține topice mai interpretabile, sugerez utilizarea de metode ce combină embeddinguri (de exemplu BERTopic sau LDA pe doc2vec) și reantrenarea sau adaptarea unui model de analiză a sentimentului la limbajul științific. Pentru cercetători, direcțiile cu potențial ridicat identificate de CAGR — contrastive learning, graph neural networks, self-supervised learning și federated learning — sunt oportunități de investiție în proiecte care pot aduce contribuții originale. Pentru evaluatori și editori, datele sugerează utilizarea unor metrici robuste (percentile, h-index, alte măsuri) în locul unei simple numărări a citărilor.

Limitări

Analiza se bazează exclusiv pe metadatele disponibile și, în general, pe titluri și abstracte, nu pe textul complet al lucrărilor; astfel, anumite nuanțe de conținut pot rămâne neobservate. Modelele standard folosite pentru topic modeling și sentiment au limite când sunt aplicate „out-of-the-box” pe limbaj academic, iar prezența datelor marcate cu ani viitori (până în 2026 în set) trebuie verificată pentru posibile erori de export sau înregistrări preliminare. În plus, redundanța terminologică identificată afectează mai multe rezultate tematice și necesită operațiuni de curățare suplimentară înainte de rafinarea concluziilor.

Propuneri de pași următori

Pentru a îmbunătăți analiza, propun normalizarea terminologică urmată de re-calcularea clusterelor, aplicarea unui model de topic pe embeddinguri pentru o separare mai clară a temelor și dezvoltarea sau adaptarea unui model de analiză a sentimentului specific limbajului academic. De asemenea, recomand verificarea înregistrărilor cu an > 2025 pentru a confirma validitatea datelor.

Dacă vrei, transform acest text cursiv într-un document gata de publicare cu secțiuni formale (Introducere, Metodologie, Rezultate, Discuții, Concluzii), tabele și figuri pregătite pentru export Word/LaTeX și, simultan, pot rula normalizarea termenilor și re-clusterizarea pe baza datelor tale — spune ce preferi și aplic imediat modificările.