IOSUD – Universitatea Politehnica Timișoara
Școala Doctorală de Studii Inginerești

Sistem de recomandare integrat pentru asistarea identificării și abordării temelor de cercetare utilizând metadate bibliografice și inteligență artificială

Teză de doctorat – Rezumat

pentru obținerea titlului științific de doctor la
Universitatea Politehnica Timișoara
în domeniul de doctorat Calculatoare și Tehnologia Informației
de către

M.Sc. Christian-Daniel Curiac
Conducător științific: Prof.univ.dr.habil.ing. Mihai Victor Micea

Septembrie 2024

Abstract: *Din când în când cercetătorii trebuie să ia în considerare noi subiecte științifice la care să lucreze dintr-o varietate de motive care provin fie din modul în care cunoștințele științifice au evoluat recent (de exemplu, dezvoltarea de noi tehnologii, teorii și metode), fie din însăși tema de cercetare actuală (de exemplu, problema de cercetare a fost deja rezolvată; nu se așteaptă noi rezultate remarcabile; tema de cercetare este blocată din cauza lipsei de idei noi sau a lipsei de resurse materiale, financiare sau umane). Descoperirea de noi teme de cercetare nu a fost niciodată atât de problematică precum astăzi, în principal din cauza dinamismului, complexității și segmentării cercetării și, de asemenea, a abundenței publicațiilor științifice care trebuie analizate. În aceste circumstanțe, cererea de instrumente automate care să-i ajute pe cercetători să descopere și să înceapă să lucreze la noi teme de cercetare promițătoare, care să răspundă atât așteptărilor cât și expertizei lor, este în creștere pronunțată. Teza de față oferă soluții pentru a acoperi această lacună prin dezvoltarea unui sistem integrat de recomandare de tip human-in-the-loop, în care informațiile contextuale, structurate, clasificate și analizate în mod adecvat, provenite dintr-un număr mare de publicații științifice, sunt folosite pentru a deriva teme de cercetare de actualitate și, mai ales, fezabile, alături de identificarea posibilelor transferuri de cunoștințe între domenii, a literaturii științifice pentru începerea cercetării sau a echipei de cercetare adecvate.*

În ultimul deceniu, Inteligența Artificială (IA) și în special Învățarea Automată (Machine Learning - ML) și-au dovedit potențialul transformator, promițând să aibă un impact revoluționar asupra economiei și societății în general. Ele au astăzi o contribuție omniprezentă în multe domenii, de la recunoașterea de imagini și a vorbirii până la diagnosticarea medicală și vehiculele autonome [1, 2]. Fiind capabilă să înțeleagă și să extragă corelații, cauzalități și modele din procesele din lumea reală, IA este din ce în ce mai probabil să înlocuiască oamenii nu numai pentru activități predictive sau repetitive, ci și pentru îndeplinirea sarcinilor cognitive complexe sau de luare a deciziilor.

Sistemele de recomandare sunt o categorie specială de aplicații IA. Ele sunt sisteme automatizate sau semi-automatizate destinate să consilieze utilizatorii în timpul luării deciziilor [3, 4]. Sistemele de recomandare au fost concepute pentru o varietate de activități din viața de zi cu

zi, cum ar fi streaming-ul de muzică sau filme ori vânzarea online asistată [5, 6, 7]. Confruntați cu o abundență de informații, utilizatorii au o sarcină din ce în ce mai dificilă oricând caută informații relevante pentru a lua decizii informate și responsabile. În acest context, tehnicile IA pot analiza riguros și în profunzime volume mari de date, iar apoi să prezinte rezultate într-o manieră inteligibilă și ușor de utilizat.

Scopul general al acestei teze este de a oferi un pachet integrat de sisteme de recomandare pentru a ajuta cercetătorii să descopere și să înceapă să lucreze la noi teme de cercetare promițătoare. Astfel, lucrarea se concentrează pe proiectarea și dezvoltarea de module de recomandare asistată (de tip human-in-the-loop), în care sunt utilizate informații contextuale reprezentând metadate bibliografice adecvat structurate și clasificate, provenite dintr-un număr mare de publicații științifice.

Descoperirea de noi teme de cercetare promițătoare a fost întotdeauna o mare provocare pentru mediul academic [8] și industrie [9]. Această activitate necesită un efort uman extins, expertiză și, nu în ultimul rând, intuiție. Procesul se bazează, de obicei, pe o parcurgere sistematică și critică a literaturii de specialitate, susținută de nevoia de a identifica modele, schimbări și tendințe într-un corp extins de cunoștințe. Ea include două etape importante, și anume identificarea ariilor de cercetare caracterizate în prezent de un deficit de cunostinte (research gaps) și formalizarea problemelor de cercetare caracteristice acestor arii. Fiecare dintre aceste etape include activități consumatoare de timp, meticuloase și uneori plictisitoare, care alternează cu luarea deciziilor în care expertiza umană joacă un rol decisiv. În urma formalizării temei de cercetare, pot fi întreprinse alte trei sarcini, și anume: investigarea posibilelor transferuri de cunoștințe care pot ajuta la rezolvarea problemei; identificarea literaturii științifice relevante pentru a ajuta la demararea evaluării stadiului actual în domeniu; și, probabil cel mai important, formarea unei echipe adecvate și eficiente pentru a desfășura activitatea de cercetare.

Odată cu apariția și adoptarea pe scară largă a tehnicilor de inteligență artificială, dezvoltarea de instrumente eficiente pentru a ajuta cercetătorii să formalizeze și să abordeze temele de cercetare relevante devine un pas firesc înainte, reliefat de următoarele motive:

- *Existența unui corp de cunoștințe din ce în ce mai mare ce trebuie analizat.* Deoarece numărul publicațiilor științifice crește exponențial, extragerea de informații utile fără instrumente automate devine din ce în ce mai dificilă. Această sarcină este suplimentar complicată de câțiva alți factori, ca de exemplu: publicațiile pot conține informații incomplete, înșelătoare sau chiar eronate; sau, accesul la unele publicații este restricționat. Este de menționat faptul că sistemele de recomandare pot elimina datele de intrare inadecvate și/sau pot prezice informațiile lipsă pe baza informațiilor existente din documente similare.

- *Fiecare cercetare are propriul ciclu de viață.* Din când în când, când temele lor devin saturate sau în declin, cercetătorii trebuie să treacă la domenii mai oportune [10].

- *Tema de cercetare aleasă trebuie să fie corelată cu formarea unei echipe capabile să o ducă la bun sfarsit, alegerea membrilor echipei putand fi descrisă ca un proces complex de optimizare* care, pe lângă descoperirea lacunelor din corpul de cunoștințe, trebuie să evalueze, să prezică și să ia în considerare cu atenție un set de constrângeri care include numărul maxim de membri ai echipei de cercetare, profilul și expertiza acestora, termenele limită, nevoile tehnice și resursele financiare disponibile [4].

- *Necesitatea corelării temelor de cercetare nou-formalizate cu tendințele actuale* și progresele în cercetare și inovare provenite din domenii emergente, care s-au dovedit a avea efecte benefice asupra întregii comunități științifice.

- *Capacitatea limitată a cercetătorului de a cuprinde și evalua întreg corpul existent de cunoștințe.* În general, este mai probabil ca cercetătorii să caute teme noi în propriile

domenii de expertiză. În acest context, o metodologie automată sau semi-automată poate ajuta la explorarea unui domeniu de cercetare mai larg, crescând astfel productivitatea și scalabilitatea procesului.

- *Interesul pentru temele de cercetare poate fi susținut de agențiile de finanțare* prin granturi sau proiecte într-un anumit domeniu de interes. În acest caz, cercetătorii trebuie să își ajusteze interesele sau temele în curs sau trebuie să găsească altele noi în domeniul definit de cererea de propuneri.

- *Necesitatea reducerii gradului de subiectivitate* în selectarea de noi teme de cercetare și echipe de cercetare, deoarece orice procedură condusă manual încorporează decizii subiective legate de incertitudinea și temerile inerente noutății domeniului ce ar trebui abordat, preocupări cu privire la timpul lung și efortul necesar pentru recalibrarea cercetătorului la o temă cu totul nouă sau îngrijorarea că rezultatele previzionate nu se vor materializa. În timp ce sistemele de recomandare pot sugera subiecte de studiu și probleme interesante, ele pot fi utilizate de asemenea pentru a identifica viitorii colaboratori care pot completa expertiza deja existentă.

- *Proiectele de cercetare tind să fie din ce în ce mai complexe* necesitând adesea o abordare multidisciplinară, cu un pronunțat caracter colaborativ.

Obiectivul principal al acestei teze este de a propune o arhitectura de sistem de recomandare integrat pentru a ajuta la identificarea și abordarea temelor de cercetare de certă actualitate și cu un impact ridicat. Arhitectura propusă implementează o metodologie de tip Human-in-the-Loop (HL) în care implicarea expertului uman este necesară pentru a supraveghea procesul de formalizare și abordare a temelor de cercetare, în principal din două motive. Primul este dat de lipsa datelor istorice pentru antrenarea modelelor IA sau ML, deoarece recomandările trebuie să fie foarte personalizate pentru un cercetător sau o echipă de cercetare care rareori desfășoară această activitate, în timp ce al doilea este legat de necesitatea de a îmbunătăți acuratețea și relevanța recomandărilor obținute, având în vedere informațiile imprecise ce caracterizează în mod inerent activitatea științifică.

Sistemul nostru integrat de recomandare se bazează pe metadatele bibliografice corespunzătoare publicațiilor științifice ce sunt colectate din baze de date bibliografice/bibliometrice (de exemplu, IEEE Xplore), acestea constituindu-se într-o sursă valoroasă și de încredere de informații legate de cercetare. În acest sens, publicațiile științifice existente sunt văzute nu doar ca rezultat direct al activităților de cercetare, ci și ca un mijloc de a înțelege tendințele de cercetare trecute, actuale și viitoare. Mai mult, prin investigarea producției științifice putem contribui la obiectivarea procesului de formare a echipei de cercetare.

Obiectivele secundare ale tezei sunt legate de proiectarea, implementarea și validarea principalelor module funcționale care constituie sistemul integrat de recomandare (sistemele de recomandare a temei de cercetare, a transferului de cunoștințe inter-domenii, și, a echipei de cercetare), având în vedere că acestea sunt menite să funcționeze și ca unități de sine stătătoare. Subiectele care sunt acoperite includ astfel achiziția și preprocesarea automată a metadatelor bibliografice, modelarea domeniilor științifice și a temelor de cercetare prin seturi de termeni cheie, evaluarea tendințelor de cercetare, transferul de cunoștințe între domenii și formarea echipelor de cercetare. Pentru a aborda subiectele de cercetare anterior menționate, au fost folosite o varietate de metode de procesare a limbajului natural (Natural Language Processing - NLP), inclusiv metode de tip entity-linking sau topic-modeling, de evaluare a similitudinii documentelor, sau de analiză a co-ocurenței termenilor. Aceste metode NLP sunt completate de tehnici de predicție și optimizare multicriterială.

Având în vedere obiectivele enunțate mai sus, contribuțiile majore oferite de această teză sunt:

- *O arhitectură de sistem integrat de recomandare de tip human-in-the-loop pentru a ajuta cercetătorii să descopere și să abordeze teme de cercetare viabile și oportune pe baza metadatelor bibliografice;*

  Analizând procesul de dezvoltare a cunoștințelor pentru un domeniu științific, este proiectat un sistem semi-automat care cuprinde patru module de recomandare (pentru tema de cercetare, pentru transferul de cunoștințe inter-domenii, pentru literatura științifică relevantă și pentru alcătuirea echipei de cercetare). Această arhitectură a fost raportată de autor în [11].

- *O metodă de evaluare a tendințelor în cercetarea științifică din metadatele bibliografice, având în vedere latența de publicare și indexare a rezultatelor cercetării;*

  Pentru a încorpora influența nefavorabilă a decalajului de timp dintre încheierea cercetării și publicarea rezultatelor acesteia asupra evaluărilor trendurilor cercetării, propunem o nouă metodologie de detecție a tendințelor. Metodologia combină metoda de predicție auto-ARIMA cu testul Mann-Kendall. Această contribuție a fost raportată în [12].

- *O metodă de identificare a subiectelor de cercetare de actualitate folosind modelarea temelor de cercetare ca seturi de termeni cheie și tehnici de predicție multivariată;*

  Reprezentând temele de cercetare ca colecții de termeni cheie, am propus o nouă abordare pentru a descoperi subiecte de cercetare cu impact din înregistrările bibliografice folosind metoda Latent Dirichlet Allocation (LDA) cuplată cu o versiune multivariată a testului Mann-Kendall. Această contribuție a fost detaliată în două dintre lucrările autorului, și anume [13] și [14].

- *O metodă de evaluare a fezabilității unei teme de cercetare folosind o metodă de tip double-threshold bazată pe co-ocurența termenilor cheie;*

  Am dezvoltat un mecanism automat pentru a identifica ariile de cercetare fezabile dar mai puțin abordate care pot fi acoperite, prin utilizarea unei proceduri cu dublu prag care filtrează temele care sunt fie dificil de studiat folosind cunoștințele existente, fie au perspective limitate de a furniza noutate. Metoda a fost subiectul lucrării [15].

- *Un sistem de recomandare a transferului de cunoștințe între domenii bazat pe conceptul de domeniu științific geamăn;*

  Teza oferă o abordare practică care utilizează metadate bibliografice pentru a identifica domeniile gemene care sunt strâns legate de un anumit domeniu științific și din care transferul de cunoștințe ar putea avea succes, precum și informațiile care ar trebui transferate.

- *Un set de date disponibil public pentru formarea echipelor de cercetare bazat pe date bibliografice/bibliometrice;*

  Setul de date constă în informații anonimizate privind expertiza tehnică și competența colaborativă a experților afiliați Universității Politehnica din Timișoara, extrase din metadatele bibliografice disponibile în IEEE Xplore pentru intervalul de timp 2010-2022. Setul de date este disponibil public în Mendeley Data [16] și este detaliat în lucrarea [17].

- *O formalizare matematică a procesului de alcătuire a echipei de cercetare ca o problemă generalizată de optimizare multi-obiectiv a acoperirii seturilor;*

  Am formulat matematic procesul de formare a echipei de cercetare ca o optimizare multi-obiectiv, personalizabilă, prin generalizarea problemei clasice multi-acoperire. Modelul nostru de optimizare este potrivit în special pentru formarea și completarea echipelor neierarhizate, dar poate fi folosit și pentru acoperirea pozițiilor non-manageriale în cadrul echipelor cu structură ierarhizată.

- *Un sistem de recomandare a echipei de cercetare care utilizează un algoritm de optimizare genetică multi-obiectiv și date bibliometrice extinse.*

  Am folosit un set extins de câmpuri de metadate bibliografice pentru a obține patru indicatori sintetici despre expertiza candidaților și abilitățile lor interpersonale. Recomandarea echipelor optime este realizată prin rezolvarea unei probleme de optimizare combinatorică multi-obiectiv folosind algoritmul genetic NSGA-II. Detaliile de proiectare și validare ale sistemului propus au fost prezentate în lucrarea [18].

În timpul studiilor de doctorat, autorul a contribuit la zece lucrări de cercetare publicate în reviste sau volumele unor conferințe internaționale, un capitol de carte și un set de date publice, după cum urmează:

1. **C.-D. Curiac**, O. Baniaș, and M. Micea, "Evaluating research trends from journal paper metadata, considering the research publication latency", Mathematics, vol. 10(2), 233, MDPI, 2022.

2. **C.-D. Curiac**, A. Doboli, and D.-I. Curiac, "Co-occurrence-based double thresholding method for research topic identification", Mathematics, vol. 10(17), 3115, MDPI, 2022.

3. **C.-D. Curiac**, and A. Doboli, "Combining informetrics and trend analysis to understand past and current directions in electronic design automation", Scientometrics, vol. 127(10), pp. 5661-5689, Springer, 2022.

4. **C.-D. Curiac**, and M. Micea, "Evaluating research trends using key term occurrences and multivariate Mann-Kendall test", in Proceedings of the International Symposium on Electronics and Telecommunications (ISETC 2022), pp. 1–4, IEEE, 2022.

5. **C.-D. Curiac**, and M. Micea, "Identifying hot information security topics using LDA and multivariate Mann-Kendall test", IEEE Access, vol. 11, pp. 18374-18384, IEEE, 2023.

6. **C.-D. Curiac**, M. Micea, T.-R. Ploscă, D.-I. Curiac, and A. Doboli "Dataset for bibliometric data-driven research team formation", Mendeley Data, doi: 10.17632r4vrvhb23h.1, 2023.

7. **C.-D. Curiac**, M. Micea, T.-R. Ploscă, D.-I. Curiac, and A. Doboli "Dataset for bibliometric data-driven research team formation: case of Politehnica University of Timisoara scholars for the interval 2010-2022", Data in Brief, vol. 53, 110275, Elsevier, 2024.

8. **C.-D. Curiac**, M. Micea, T.-R. Ploscă, D.-I. Curiac, and A. Doboli "Optimized interdisciplinary research team formation using a genetic algorithm and extended bibliometric data". *[under review]*

9. **C.-D. Curiac**, M. Micea, T.-R. Ploscă, D.-I. Curiac, S. Doboli and A. Doboli "Towards automating new research problem framing and exploration based on symbolic-numerical

knowledge extracted from bibliometric data", in Bibliometrics - An Essential Methodological Tool for Research Projects. IntechOpen, London, UK, 2024.

10. T. Andreica, **C.-D. Curiac**, C. Jichici and B. Groza, "Android head units vs. in-vehicle ECUs: performance assessment for deploying in-vehicle intrusion detection systems for the CAN bus", IEEE Access, vol. 10, pp. 95161-95178, IEEE, 2022.

11. M.D. Baciu, E.A. Capota, C.S. Stângaciu, **C.-D. Curiac**, and M. Micea, "Multi-core time-triggered OCBP-based scheduling for mixed criticality periodic task systems", in Proceedings of the International Symposium on Electronics and Tele-communications (ISETC 2022), pp. 1–4, IEEE, 2022.

12. T.-R. Ploscă, **C.-D. Curiac**, and D.-I. Curiac. "Investigating semantic differences in user-generated content by cross-domain sentiment analysis means", Applied Sciences, vol. 14(6), 2421, MDPI, 2024.

Primele nouă dintre aceste lucrări reprezintă pilonii principali ai tezei, în timp ce restul abordează subiecte de învățare automată și de procesare în timp real.

În cele ce urmează, este prezentat un scurt rezumat al principalelor capitole din teză.

Capitolul 2 oferă o analiză critică a stadiului actual în domeniul sistemelor de recomandare pentru sprijinirea cercetării, subliniind metodele și tehnicile existente pentru a recomanda conținut științific, citări, teme de cercetare, precum și membri ai echipei pentru proiectele de cercetare.

Din această analiză merită menționate următoarele concluzii: (a) toate lucrările menționate folosesc abordări text-mining pentru a investiga seturi de publicații științifice, astfel de tehnici NLP dovedind rezultate promițătoare; (b) cercetarea în domeniu este încă la început, nereușind să ofere sisteme de recomandare integrate care să-i ajute în mod adecvat pe cercetători atunci când își încep sau își desfășoară cercetarea; și, (c) cercetarea existentă este limitată și fragmentată, fiind îndreptată doar către patru obiective (propunerea de conținut științific personalizat, sugerarea de citări care să însoțească o temă de cercetare, descoperirea punctelor fierbinți de cercetare pe baza analizei tendințelor, și, formarea echipelor de cercetare) neglijând în același timp aspecte importante precum identificarea temelor de cercetare viabile și actuale sau transferurile de cunoștințe între domenii. Din perspectiva noastră, un sistem integrat de recomandare human-in-the-loop pentru asistarea cercetătorilor în descoperirea, formalizarea și începerea de noi teme personalizate de cercetare poate accelera procesul de cercetare și oferă un nivel sporit de obiectivitate. În cazul nostru, necesitatea ca un expert uman să supravegheze formalizarea și selecția temei de cercetare și activitățile aferente acesteia este determinată de două motive. Primul este dat de lipsa datelor pentru a antrena modele de învățare automată sau inteligență artificială, deoarece recomandările trebuie personalizate pentru un cercetător sau o echipă de cercetare care rareori desfășoară această activitate, în timp ce al doilea este legat de necesitatea îmbunătățirii preciziei și relevanța recomandărilor.

Capitolul 3 prezintă arhitectura propusă pentru sistemul integrat de recomandare menit să ajute cercetătorii în identificarea și explorarea subiectelor de cercetare relevante folosind informații extrase din înregistrările bibliografice corespunzătoare publicațiilor științifice. Acestă arhitectură încorporează patru module de recomandare care, pe baza metadatelor bibliografice, asistă cercetătorii în găsirea de noi teme de cercetare, în identificarea cunoștințele care sunt potrivite pentru a fi transferate din alte domenii științifice, în găsirea unui set de publicații relevante cu care să înceapă analiza stadiului actual în domeniu, și în formarea unei echipe adecvată de experți care să finalizeze tema. Efortul nostru s-a bazat pe modul în care pot apărea cunoștințe noi într-un anumit domeniu științific, subliniind că ideile de cercetare semnificative

pot fi derivate din combinații adecvate de cunoștințe deja existente în domeniu, pot proveni din domenii strâns legate și emergente sau, mai rar, pot rezulta din idei novatoare apărute aparent din senin (sudden insights). Capitolul încapsulează arhitectura prezentată în [11].

Metodologia noastră urmărește automatizarea întregului proces care precede cercetarea științifică și începe cu formularea de noi teme de cercetare fezabile și de actualitate, continuă cu recomandarea literaturii științifice relevante și se încheie cu formarea de echipe pentru implementarea temei. Abordarea propusă este definită printr-o procedură semi-supravegheată, întrucât expertul uman, în cazul nostru cercetătorul, intervine pentru a canaliza procesul în funcție de așteptările și expertiza sa.

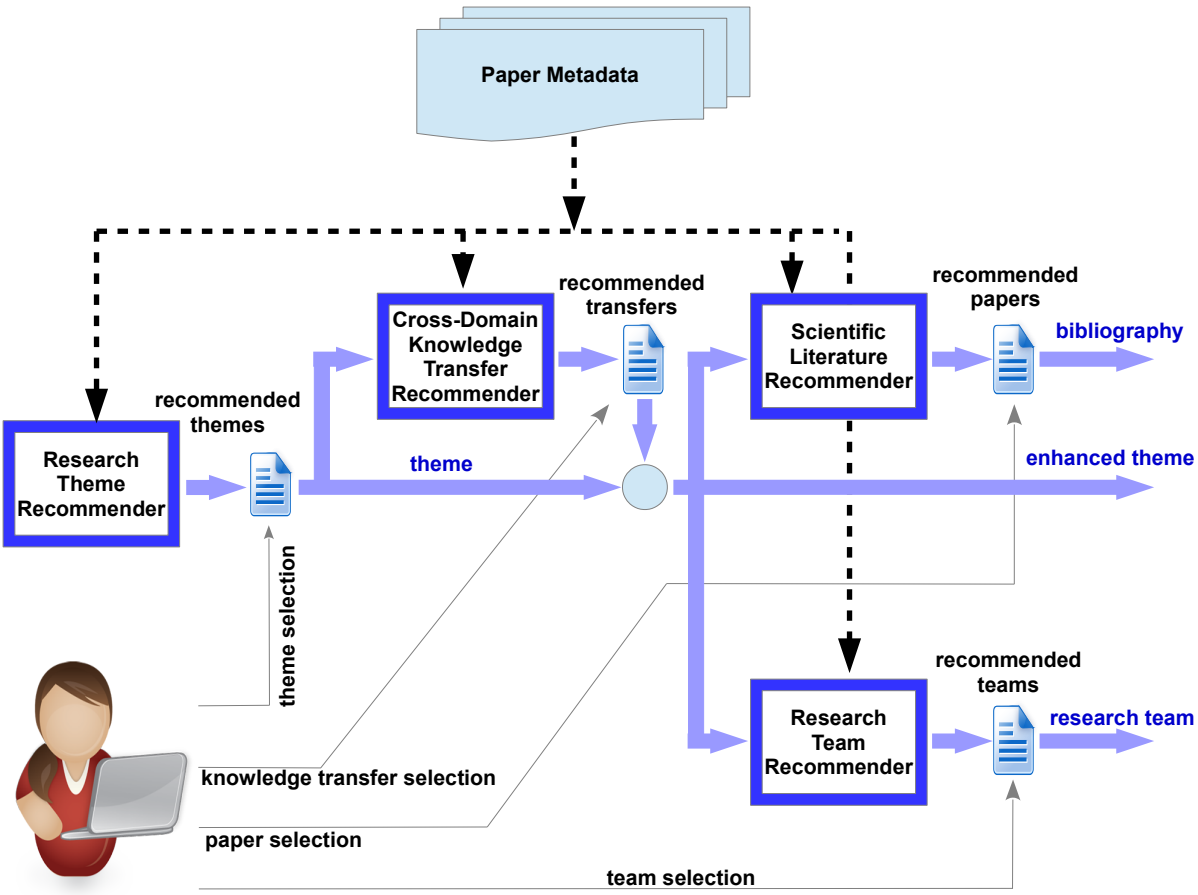

Figura 1: Arhitectura sistemului integrat de recomandare

Intrarea acestei proceduri complexe de recomandare este reprezentată de un set bogat de date care conține metadate ale lucrărilor științifice (înregistrări bibliografice care conțin informații legate de publicații, inclusiv numele și afilierile autorilor, titlul, cuvinte cheie și abstractul lucrării) care pot rezuma în mod eficient cercetarea în domeniu și domenii conexe sau emergente și pot urmări de asemenea profilurile de publicare ale cercetătorilor. Pentru a obține setul de date necesar, putem extrage înregistrări din baze de date bibliometrice influente precum Clarivate Web of Science, Scopus, PubMed sau IEEE Xplore. Pentru studiile de caz și experimentele noastre, deoarece ne-am îndreptat atenția către teme de cercetare din domeniul tehnologiei informației și al ingineriei electrice și electronice, am selectat IEEE Xplore ca sursă de date bibliometrice.

Arhitectura propusă conține patru module de recomandare care pot acționa fie interconectate

ca în Figura 1, fie ca sisteme de recomandare autonome:

I. Research theme recommender (sistem de recomandare a temelor de cercetare),

II. Cross-domain knowledge transfer recommender (sistem de recomandare a transferului de cunoștințe între domenii),

III. Scientific literature recommender (sistem de recomandare a literaturii științifice de specialitate), și

IV. Research team recommender (sistem de recomandare a echipei de cercetare).

Research theme recommender. Acest sistem de recomandare human-in-the-loop își propune să ofere o listă de teme de cercetare actuale și fezabile prin investigarea stării cercetării în domeniu și prin judecarea oportunității și viabilității temelor prin efectuarea de analize statistice și de trenduri. În acest context, o colecție de metadate relevante ale publicațiilor este utilizată pentru a identifica o listă extinsă de termeni specifici domeniului, pentru a descoperi lacunele de cercetare existente în cadrul domeniului și, de asemenea, pentru a evalua oportunitatea și importanța rezolvarii problemelor care stau în spatele acestor lacune în cercetare.

Cross-domain knowledge transfer recommender. Acest sistem de recomandare este menit să identifice posibile surse pentru transferuri relevante de cunoștințe care pot ajuta la rezolvarea temei de cercetare recomandate. Folosind evaluarea similarității documentelor și tehnici de topic-modeling, explorăm domenii gemene și emergente pentru a găsi cunoștințe care și-au dovedit deja eficiența, capabile să fie legate de tema de cercetare propusă.

Scientific literature recommender. Pentru a ajuta cercetătorii să stabilească un punct de plecare potrivit de unde poate începe analiza literaturii de specialitate din domeniu, ne îndreptăm căutarea în două direcții: o explorare în domeniu pentru a găsi lucrări fundamentale referitoare la tema de cercetare; și, explorări de domenii gemene/emergente pentru a găsi lucrări relevante referitoare la cunoștințele pe care intenționăm să le transferăm.

Research team recommender. Analizând corpusul de metadate bibliografice pentru a extrage indicatori despre expertiza și abilitățile de lucru în echipă asociate fiecăruia dintre cercetătorii disponibili, se propune un set de echipe care pot realiza tema de cercetare specificată prin rezolvarea unei probleme complexe de optimizare multi-obiectiv a formării echipei.

La sfârșitul procesului de recomandare propus, este oferită o listă de posibile teme de cercetare actuale și, mai ales, fezabile, însoțită de oportunități de transfer de cunoștințe, propuneri de bibliografie științifică și recomandări ale echipei de cercetare.

Capitolul 4 descrie procedura de achiziție și preprocesare a datelor. Ne bazăm pe interfețele de tip API (Application Programming Interface) pentru a colecta automat înregistrările de metadate ale publicațiilor corespunzătoare revistelor științifice de top sau conferințelor anuale pentru a rezuma eficient cercetarea într-un anumit domeniu. Folosind o tehnică adecvată de tip entity-linking (de exemplu, TagMe), câmpurile conţinând titlul, cuvintele cheie și abstractul sunt transformate în liste de termeni cheie relevanți pentru a reprezenta esența conținutului publicației. O parte din rezultatele acestui capitol au fost raportate în [17] și a fost oferit ca set de date public pe Mendeley Data [16].

Capitolul 5 este dedicat prezentării detaliate a modulului semi-automat de recomandare a temei de cercetare, capabil să sugereze subiecte de cercetare noi și de mare impact. Metodologia noastră e de tip human-in-the-loop. Ea începe cu identificarea setului de termeni cheie care caracterizează domeniul științific și grupează acești termeni cheie în teme de cercetare. Ulterior, temele de cercetare sunt investigate din punct de vedere al oportunității și fezabilității lor prin folosirea unei analize statistice complexe ce include evaluarea multivariabilă a trendurilor. Capitolul încapsulează metodele și rezultatele prezentate în lucrările [12, 13, 14, 15].

Urmând această strategie, am proiectat și construit un sistem de recomandare având arhitectura prezentată în Figura 2.

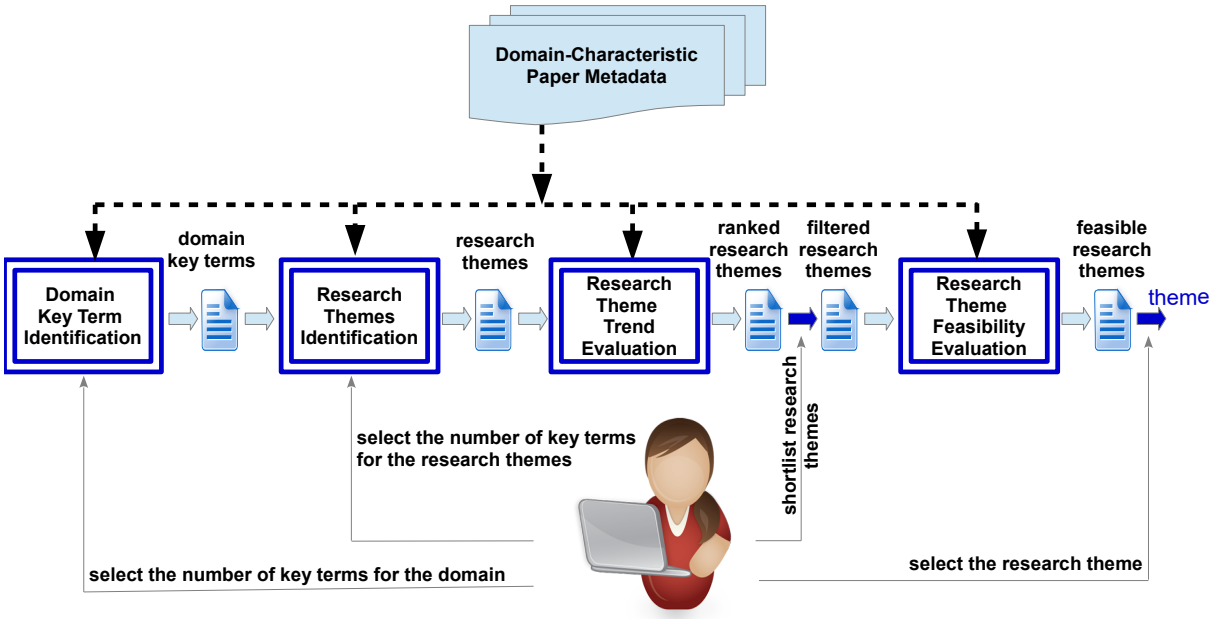

Figura 2: Arhitectura sistemului de recomandare a temelor de cercetare

Acest sistem de recomandare are ca intrare un corpus C suficient de mare de metadate bibliografice și oferă o listă de teme de cercetare de actualitate și, de asemenea, fezabile, care vor fi supuse examinării critice a utilizatorului pentru selectarea celei mai bune alternative. Corpusul C de metadate prelucrate trebuie să îndeplinească următoarele cerințe: (i) să acopere complet și, dacă este posibil, uniform întregul domeniu D, fără a lăsa neacoperit niciun subdomeniu sau zonă științifică din cadrul domeniului; (ii) să aibă o acoperire continuă în timp a domeniului pe o durată de cel puțin zece ani pentru a identifica în mod eficient tendințele cercetării; și, (iii) să includă numai materiale științifice recenzate pentru a certifica că rezultatele cercetărilor publicate sunt originale, logice, semnificative și amănunțite. Pentru aceasta, utilizatorul trebuie să selecteze periodicele emblematice ale domeniului (de exemplu, reviste de renume și conferințe anuale) pe baza reputației și recunoașterii lor în comunitatea științifică, reflectată de indici bibliometrici excepționali (de exemplu factorul de impact al jurnalului Clarivate, CiteScore al Elsevier). Spre exemplificare, în cazul investigării temelor de cercetare din domeniul Electronic Design Automation, putem lua în considerare reviste precum IEEE Transactions on Computer Aided Design of Integrated Circuits & Systems (revista emblematică a IEEE Council on Electronic Design Automation) și ACM Transactions on Design Automation for Electronic Systems (revista emblematică a ACM Special Interest Group on Design Automation) și conferințe anuale prestigioase precum Design Automation (DAC); Design, Automation, Test in Europe (DATE); Asia and South Pacific Design Automation (ASPDAC); și, International Conference on Computer-Aided Design (ICCAD).

Acest sistem de recomandare constă din patru blocuri funcționale care sunt executate succesiv:

1. *Domain Key Term Identification (identificarea termenilor cheie din domeniu):* furnizează o listă cuprinzătoare de termeni cheie pentru a modela cu acuratețe starea actuală a domeniului științific D prin evaluarea frecvențelor termenilor cheie în metadatele bibliografice din ultimii unu până la trei ani;

2. *Research Themes Identification (identificarea temelor de cercetare):* extrage temele de cercetare care caracterizează domeniul științific D utilizând metoda LDA pe același corpus de documente folosit de blocul anterior;

3. *Research Theme Trend Evaluation (evaluarea tendinței temei de cercetare):* investighează actualitatea fiecărei teme de cercetare prin evaluarea trendului acesteia cu o variantă multivariabilă adecvată a testului clasic Mann-Kendall. În cazul în care latența de publicare/indexare corespunzătoare lucrărilor din reviste sau conferințe nu poate fi ignorată, pentru a calcula tendințe mai precise a fost concepută o metodă nouă care combină metodele auto-ARIMA și testul Mann-Kendall în forma multivariabilă;

4. *Research Theme Feasibility Evaluation (evaluarea fezabilității temei de cercetare)*: examinează fiecare temă de cercetare în ceea ce privește noutatea și succesul prezumat și o clasifică ca fezabilă sau nu folosind o metodă cu dublu prag.

Capitolul 6 descrie modulul de recomandare conceput pentru a identifica posibilele transferuri de cunoștințe între domenii, cu precadere din domenii științifice similare/gemene sau emergente. Acest capitol oferă o metodologie practică în care metadatele bibliografice sunt folosite pentru a descoperi domeniile gemene legate de domeniul țintă de unde transferul poate fi eficient și, de asemenea, cunoștințele din domeniile gemene și emergente care urmează să fie transferate și personalizate. În acest sens, s-au folosit tehnici de analiză a similarității documentelor și de tip topic-modeling.

Sistemul de recomandare propus pentru transferul de cunoștințe preia rezumatele procesate ale lucrărilor publicate și setul de termeni cheie care modelează tema de cercetare și oferă un set de transferuri de cunoștințe (termeni cheie) din domenii gemene sau emergente. Arhitectura sa este prezentată în Figura 3 și conține patru pași funcționali.

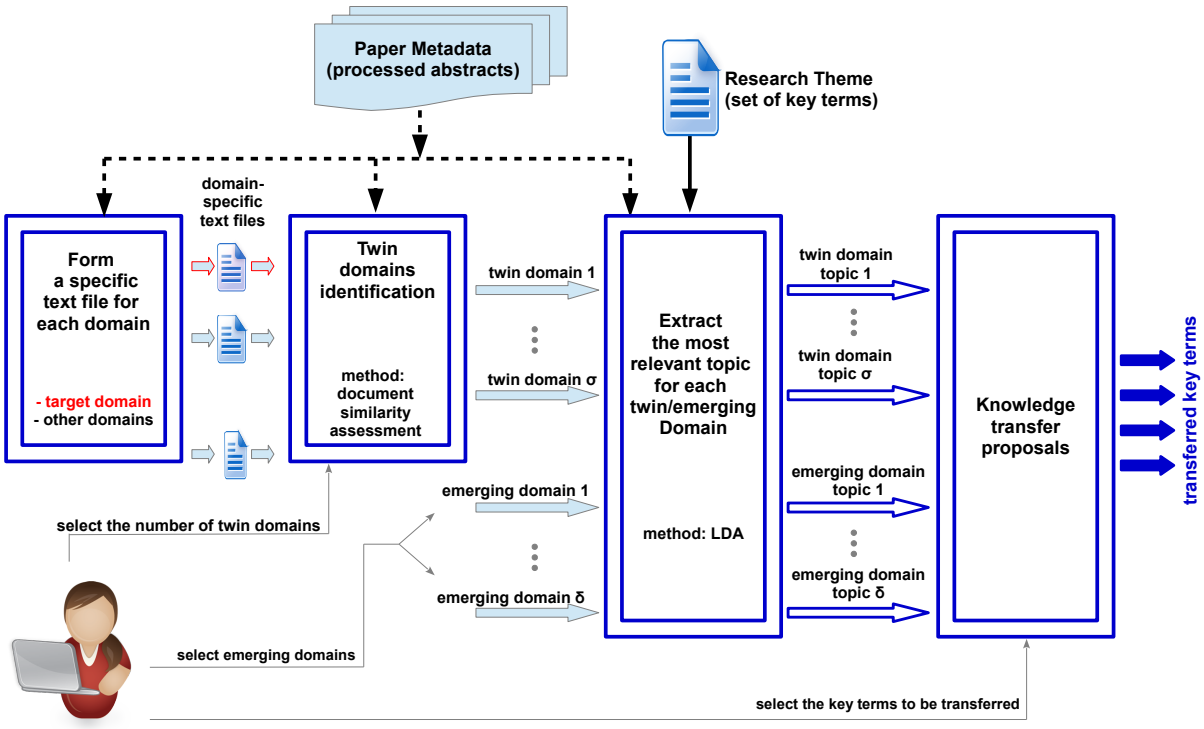

Figura 3: Arhitectura sistemului de recomandare a transferului de cunoștințe între domenii

Pasul 1 (Construirea unui fișier text caracteristic domeniului pentru fiecare domeniu științific): Având scopul de a reprezenta cât mai exact posibil corpul de cunoștințe al fiecărui domeniu științific, concatenăm într-un singur document text rezumatele procesate corespunzătoare tuturor lucrărilor publicate în cadrul domeniului investigat. Prin aceasta, toate informațiile semnificative extrase din titlurile publicațiilor, cuvintele cheie și abstracturi sunt rezumate utilizând o tehnica de tip entity-linking sub forma unei liste de termeni cheie care va fi folosită pentru a identifica domeniile științifice gemene ale unui anumit domeniu țintă.

Pasul 2 (Identificarea domeniilor gemene): Pentru a identifica domeniile sursă care, datorită apropierii lor de domeniul țintă (în ceea ce privește subiectele, metodele și materialele), sunt plauzibile să exporte cunoștințe valoroase, examinăm gradul de similitudine dintre perechile de documente ce corespund domeniului țintă și domeniului sursă. Acest pas urmează o procedură clasică de evaluare a similitudinii documentelor care utilizează term frequency–inverse document frequency (*tf-idf*) corespunzător modelului vectorial și o metrică de evaluare a similarității (cosine similarity measure) [19]. Această metodă include două etape:

a) vectorizarea corpus-ului de documente utilizand *tf-idf* [20] pentru toți termenii (entitățile) cheie întâlniți. Indicatorul *tf-idf* este menit să caracterizeze importanța termenilor pentru un document dintr-o colecție de documente.

b) să evalueze asemănările dintre perechile de fișiere text caracteristice domeniului țintă și respectiv domeniului sursă utilizând metrica cosine similarity [21].

În cele din urmă, le vom desemna ca domenii gemene pe cele având cei mai mari indici de similaritate.

Pasul 3 (Clusterizarea corpului de cunoștințe în fiecare domeniu sursă folosind metoda LDA): Pentru a identifica contextul în care cunoștințele presupus-legate de tema științifică dată RT apar în domenii gemene sau emergente, utilizăm procedura Latent Dirichlet Allocation. Astfel, pentru fiecare dintre domeniile gemene sau emergente identificate, aplicăm LDA pentru a clasifica termenii domeniului geamăn/emergent în $4-8$ grupuri și identificăm topicele care conțin cel mai mare număr de termeni cheie KT_j care descriu RT. Ulterior, calculăm co-ocurența termenilor cheie dintre termenii definitorii KT_j ai RT și alți termeni care se află în același topic și reținem termenii din domenii gemene/emergente care au valori de co-ocurență peste un prag ales pentru a fi analizați pentru un posibil transfer de cunoștințe. Este de menționat că atunci când analizăm domeniile gemene/emergente suntem interesați să identificăm domeniile în care cercetarea este mai avansată și poate fi o sursă de transferuri valoroase de cunoștințe. Astfel de zone avansate sunt caracterizate în matricea de co-ocurență $\mathcal{M}$ cu un scor mare. În acest fel, putem transfera cunoștințe de mare impact din domenii conexe către RT.

Pasul 4 (Recomandarea transferului de cunoștințe): în acest pas, posibilele transferuri de cunoștințe sunt prezentate utilizatorului sub formă de seturi de termeni cheie care pot însoți setul existent de termeni cheie KT_j descriind RT.

Capitolul 7 prezintă modulul de recomandare a echipei de cercetare. Pentru a formaliza matematic procesul de formare a echipei, propunem un model multicriterial generalizat de optimizare a acoperirii seturilor care poate face față unei mari varietăți de obiective și constrângeri asociate echipei. Utilizând un set extins de date bibliografice și bibliometrice, evaluăm expertiza tehnică și abilitățile de colaborare ale fiecărui candidat pe baza a patru descriptori proiectați cu atenție și rezolvăm problema rezultată utilizând algoritmul evolutiv elitist NSGA-II.

Bazat pe evaluările candidaților extrase din metadatele bibliografice, propunem o metodologie generală pentru formarea echipelor de cercetare neierarhizate. Diagrama acestei metodologii care implementează un sistem de recomandare human-in-the-loop este prezentată în Figura 4. După cum putem observa, sistemul de recomandare are următorul set de intrări: (i) un corpus de metadate bibliografice atent ales; (ii) specificațiile proiectului de cercetare care trebuie

îndeplinite; și, (iii) detalii privind contextul organizațional în care va funcționa echipa de cercetare rezultată.

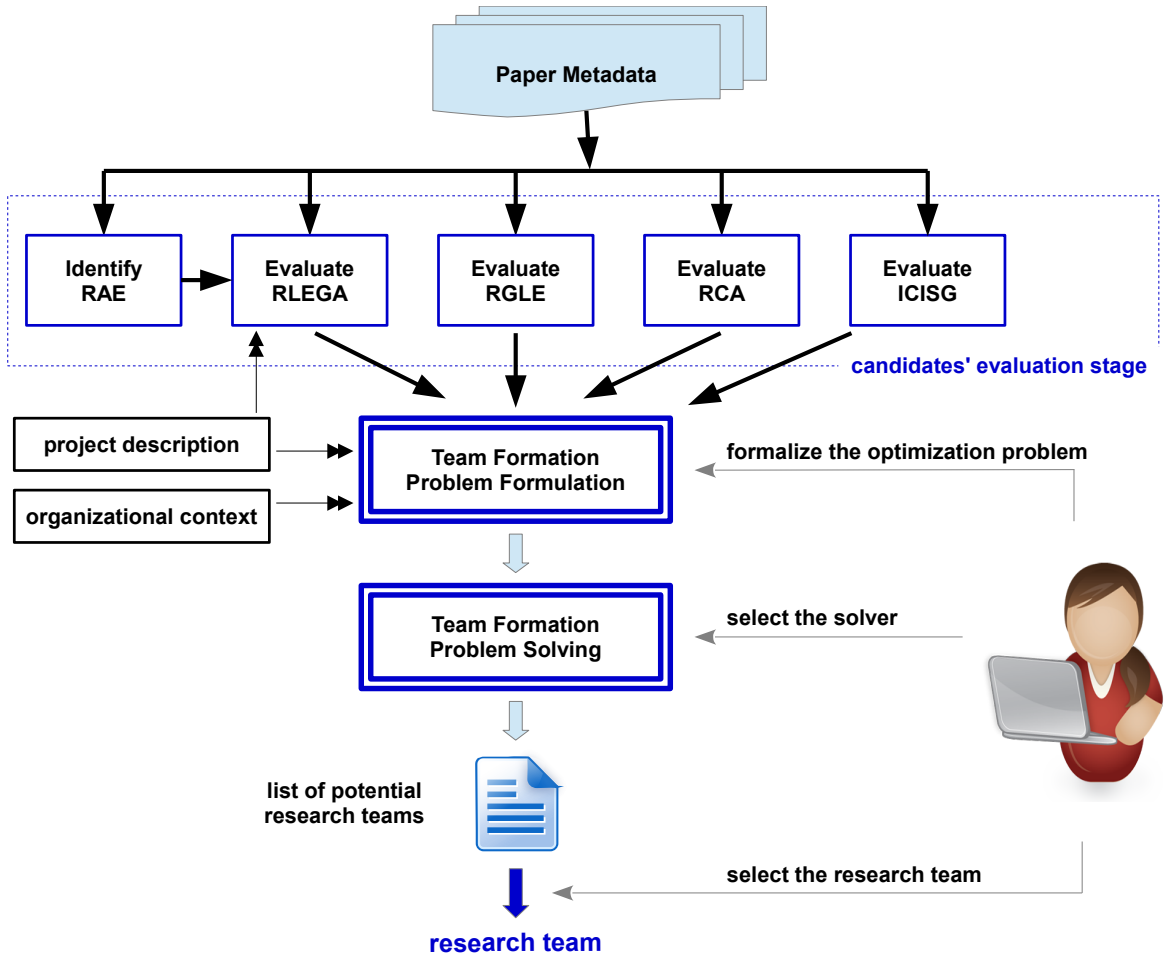

Figura 4: Arhitectura sistemului de recomandare a echipei de cercetare

Prima etapă a acestei proceduri se concentrează pe preprocesarea metadatelor bibliografice și pe evaluarea candidaților. Ea oferă, pe lângă ariile de expertiză ale cercetătorului (RAE), un set de indicatori privind abilitățile tehnice și nontehnice ale candidaților: expertiza generală a cercetătorului (RGE), expertiza cercetătorului într-un domeniu specificat (RLEGA), abilitatea de colaborare a cercetătorului (RCA), abilitatea de colaborare interpersonală în cadrul unor grupuri de dimensiune specificată (ICISG).

În următoarea etapă a metodologiei noastre, formalizarea problemei de formare a echipei se realizează luând în considerare caracteristicile solicitate candidaților, informațiile disponibile despre proiectul de cercetare și contextul organizațional (buget, interacțiune cu alte proiecte de cercetare, locație, infrastructură de cercetare, etc.). Ca rezultat, se obține o problemă de optimizare combinatorică multi-obiectiv specifică proiectului, care poate fi ulterior remodelată sau chiar simplificată pentru a îndeplini cerințele unei metode alese de rezolvare a problemelor.

Lista echipelor sugerate este oferită inițiatorului echipei, care poate alege componența sa favorită. Dacă procesul produce rezultate inadecvate (de exemplu, echipe de cercetare conflictuale), inițiatorul poate reporni secvența de formare a echipei făcând modificări corespunzătoare în etapele precedente (de exemplu, încercând să colecteze informații noi, sau să modifice formularea

problemei de optimizare prin remodelarea funcţiilor obiectiv sau a constrângerilor).

După cum se poate observa din Figura 4, metodologia propusă de noi este una asistată, unde inițiatorul echipei joacă un rol decisiv nu doar în formularea problemei de optimizare, ci și în alegerea metodei de rezolvare, sau în alegerea celei mai potrivite echipe pentru îndeplinirea obiectivului de cercetare.

Capitolul 8 oferă un rezumat al tezei, urmat de concluzii, o scurtă discuție despre limitările metodelor propuse și despre direcțiile de cercetare viitoare.

Întrucât, în opinia noastră, această lucrare este printre primele care abordează problema formalizării și abordării de noi teme de cercetare cu ajutorul metadatelor bibliografice, aria de cercetare este larg deschisă. Pentru a îmbunătăți și mai mult sistemul de recomandare propus, merită menționate cinci direcții de cercetare: (a) automatizarea selecției și reglajul fin al parametrilor utilizați de tehnicile IA utilizate; (b) includerea de noi surse de informații privind cercetarea științifică (de exemplu, baze de date care conțin proiecte de cercetare precum CORDIS sau colecții de proiecte software precum GitHub); (c) tratarea existenței publicațiilor științifice false ori nerelevante; (d) validarea sistemului integrat de recomandare propus pe alte baze de date bibliografice/bibliometrice relevante, cum ar fi PubMed, Scopus, Web of Science sau Scopus, și analizarea modului în care informațiile obținute de la diverse surse bibliografice pot spori acuratețea abordării propuse; (e) proiectarea unui sistem eficient și mai orientat către client pentru recomandările literaturii științifice, această problemă fiind abordată doar tangenţial în cadrul acestei teze.

BIBLIOGRAFIE

[1] G. Wisskirchen, B. Biacabe, U. Bormann, A. Muntz, G. Niehaus, G. Soler, and B. von Brauchitsch, "Artificial intelligence and robotics and their impact on the workplace," *IBA Global Employment Institute*, vol. 11, no. 5, pp. 49–67, 2017.

[2] J. Andreu-Perez, F. Deligianni, D. Ravi, and G.-Z. Yang, "Artificial intelligence and robotics," https://arxiv.org/ftp/arxiv/papers/1803/1803.10813.pdf, 2016, accessed: 2024-04-25.

[3] J. Beel, B. Gipp, S. Langer, and C. Breitinger, "Research paper recommender systems: A literature survey," *International Journal on Digital Libraries, Springer*, vol. 17, pp. 305–338, 2016.

[4] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 734–749, 2005.

[5] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez, "Recommender systems: Survey," *Knowledge-Based Systems*, vol. 46, pp. 109–132, 2013.

[6] K. Wei, J. Huang, and S. Fu, "A survey of e-commerce recommender systems," in *Proc. of the International Conference on Service Systems and Service Management.* IEEE, Chengdu, China, Jul. 2007, pp. 1–5.

[7] M. Schedl, P. Knees, and F. Gouyon, "New paths in music recommender systems research," in *Proc. of the Eleventh ACM Conference on Recommender Systems.* ACM, Como, Italy, 2017, p. 392–393.

[8] H. Lee and P. Kang, "Identifying core topics in technology and innovation management studies: A topic model approach," *The Journal of Technology Transfer*, vol. 43, pp. 1291–1317, 2018.

[9] H. Small, K. Boyack, and R. Klavans, "Identifying emerging topics in science and technology," *Research Policy*, vol. 43, no. 8, pp. 1450–1467, 2014.

[10] T. Kuhn, "The structure of scientific revolutions," *University of Chicago press*, 1962.

[11] C.-D. Curiac, M. Micea, T.-R. Plosca, D.-I. Curiac, S. Doboli, and A. Doboli, "Towards automating new research problem framing and exploration based on symbolic-numerical knowledge extracted from bibliometric data," in *Bibliometrics - An Essential Methodological Tool for Research Projects*. IntechOpen, London, UK, 2024, *[accepted for publication]*.

[12] C.-D. Curiac, O. Banias, and M. Micea, "Evaluating research trends from journal paper metadata, considering the research publication latency," *Mathematics*, vol. 10, no. 2, p. 233, 2022.

[13] C.-D. Curiac and M. Micea, "Evaluating research trends using key term occurrences and multivariate Mann-Kendall test," in *2022 International Symposium on Electronics and Telecommunications (ISETC)*. IEEE, Timișoara, Romania, 2022, pp. 1–4.

[14] C.-D. Curiac and M. V. Micea, "Identifying hot information security topics using LDA and multivariate Mann-Kendall test," *IEEE Access*, vol. 11, pp. 18 374–18 384, 2023.

[15] C.-D. Curiac, A. Doboli, and D.-I. Curiac, "Co-occurrence-based double thresholding method for research topic identification," *Mathematics*, vol. 10, no. 17, p. 3115, 2022.

[16] C.-D. Curiac, M. Micea, T.-R. Plosca, D.-I. Curiac, and A. Doboli, "Dataset for bibliometric data-driven research team formation," *Mendeley Data*, 2023, doi: 10.17632/r4vrvhb23h.1.

[17] C.-D. Curiac, M. Micea, T.-R. Plosca, D.-I. Curiac, and A. Doboli, "Dataset for bibliometric data-driven research team formation: Case of Politehnica University of Timisoara scholars for the interval 2010-2022," *Data In Brief*, vol. 53, p. 110275, 2024.

[18] C.-D. Curiac, M. Micea, T.-R. Plosca, D.-I. Curiac, and A. Doboli, "Optimized interdisciplinary research team formation using a genetic algorithm and publication metadata records," *[under review]*.

[19] D. Sailaja, M. Kishore, B. Jyothi, and N. Prasad, "An overview of pre-processing text clustering methods," *International Journal of Computer Science and Information Technologies*, vol. 6, no. 3, pp. 3119–24, 2015.

[20] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.

[21] G. Salton, *Automatic text processing: The transformation, analysis, and retrieval of*. Addison-Wesley, Boston, USA, 1989, vol. 169.