UNIVERSITATEA BABEŞ-BOLYAI, CLUJ-NAPOCA, ROMÂNIA, FACULTATEA DE MATEMATICĂ ŞI INFORMATICĂ

Optimizarea Sistemelor de Recomandare: Strategii pentru Imbunătăţirea Performanţei si Adaptabilităţii

Sumarul tezei de doctorat

Student doctorand: Mara Deac-Petruşel
Coordonator ştiinţific: Professor PhD Anca Andreica

2024

Cuvinte cheie: Sisteme de recomandare, Filtrare Colaborativă, Metoda celor mai apropiaţi k vecini, Predicţia Recenziilor, Învățare automată

Rezumat

Într-o lume dominată de abundența datelor și experiențe personalizate, Sistemele de Recomandare joacă un rol esențial în ghidarea proceselor decizionale. Printr-o investigație meticuloasă a diferitelor metodologii și formularea unei noi măsuri de similitudine, obiectivul principal al acestei teze este de a înbunătăţii calitatea recomandărilor.

Un prim focus al tezei este analizarea tehnicilor de filtrare colaborativă bazate pe memorie, cu un accent deosebit pe rolul crucial al măsurilor de similaritate. Experimente ample pe diverse seturi de date dezvăluie măsuri de similaritate optime pentru diferite contexte.

În continuare, este introdusă o nouă măsură de similaritate bazată pe sentiment, numită Atracție-Relevanță-Popularitate (ARP), având ca scop îmbunătățirea filtrării colaborative prin valorificarea recenziilor textuale. ARP înlocuiește evaluările numerice cu scoruri de sentiment derivate din lexicoane de analiză a sentimentelor, rezultând o acuratețe îmbunătățită a recomandărilor. În plus, este propus un cadru robust de validare pentru ARP, vizând revoluționarea procesului de dezvoltare și evaluare a noilor măsuri de similaritate.

Mai mult, au fost concepute trei abordări cu scopul de a optimiza tehnicile de recomandare. Prima abordare utilizează tehnici de analiză a sentimentelor în combinație cu filtrarea colaborativă, demonstrând îmbunătățiri semnificative în acuratețea și calitatea recomandărilor. A doua abordare introduce o tehnică de filtrare colaborativă KNN bazată pe lexicoane, demonstrând succes în sarcinile de recomandare cu seturi de date care conțin recenzii textuale ale utilizatorilor. Ultima secțiune prezintă un sistem de recomandare nesupervizat, adaptat pentru cititorii New York Times, incorporând clustering-ul K-Means pentru a defini clusterele de articole și a genera recomandări personalizate.

Această teză oferă perspective valoroase asupra diferitelor aspecte ale sistemelor de recomandare, inclusiv studiul și dezvoltarea măsurilor de similaritate, fuziunea cu tehnicile de analiză a sentimentelor și surprinzătoarea abordare nesupervizată a sistemelor de recomandare. Concluziile contribuie la avansarea cercetării în domeniul sistemelor de recomandare, stabilind o fundație solidă pentru explorări și inovații viitoare în acest domeniu.

Lista Publicațiilor

- Petruşel, Mara and Limboi, Sergiu-George. **A restaurants recommendation system: Improving rating predictions using sentiment analysis.** In 2019 21st International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), pages 190-197, 2019 IEEE [19] (Conference Category C - 2 points).

- Deac-Petruşel, Mara, **A comparative analysis of similarity measures in memory-based collaborative filtering.** In Artificial Intelligence and Soft Computing: 19th International Conference, ICAISC 2020, Zakopane, Poland, October 12-14, 2020, Proceedings, Part II 19, pages 140-151, Springer International Publishing, 2020 [6] (Conference Category C - 2 points)

- Deac-Petruşel, Mara and Limboi, Sergiu. **A sentiment-based similarity model for recommendation systems.** In 2020 22nd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), pages 224-230, 2020. IEEE [8] (Conference Category D - 1 point)

- Limboi, Sergiu and Deac-Petruşel, Mara. **A Validation Framework for ARP Similarity Measure**. In 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), pages 1266-1271, 2021 IEEE [15] (Conference Category C - 2 points)

- Deac-Petruşel, Mara. **A Lexicon-based Collaborative Filtering Approach for Recommendation Systems.** In Proceedings of the 14th International Conference on Agents and Artificial Intelligence (ICAART), pages 203-210, 2022, SCITEPRESS [7] (Conference Category B - 4 points)

- Petruşel, Mara. **An Unsupervised Topic-driven New York Times Recommendation System.** In 2022 International Conference on Innovations in Intelligent Systems and Applications (INISTA), pages 1-6, 2022, IEEE [18] (Conference Category C - 2 points)

Total: 13 puncte.

Cuprins

Capitolul 1

Introducere

1.1 Motivație

Într-o eră caracterizată de o creștere fără precedent a disponibilității datelor, domeniul Sistemelor de Recomandare (SR) se evidențiază ca o piatră de temelie a inovației tehnologice. Esența acestor sisteme nu constă doar în capacitatea lor de a sugera produse, servicii sau conținut, ci în puterea lor de a influența deciziile, de a stimula creșterea economică și de a spori satisfacția utilizatorilor.

Impactul profund al Sistemelor de Recomandare se resimte în diverse domenii, de la platformele de comerț electronic care personalizează experiențele de cumpărare în funcție de preferințele individuale, până la serviciile de streaming de conținut care creează liste de redare adaptate gusturilor. În peisajul complex al deciziilor moderne, semnificația acestor sisteme depășește confortul; ele captează interacțiunea complexă dintre preferințele utilizatorilor, recuperarea informațiilor și progresele tehnologice.

Motivația pentru a studia sistemele de recomandare derivă din natura lor dinamică și în continuă evoluție. Într-un peisaj digital caracterizat de un flux neîncetat de date, abilitatea de a extrage informații semnificative și de a îmbunătăți acuratețea recomandărilor nu este doar o preocupare academică, ci un răspuns la nevoile în schimbare ale societății.

Mai mult, pe măsură ce ecosistemele digitale devin din ce în ce mai complexe și interconectate, rolul Sistemelor de Recomandare devine și mai critic. Ele servesc drept instrumente vitale pentru gestionarea supraîncărcării informaționale, ajutând utilizatorii să navigheze prin cantități vaste de date prin prezentarea celor mai relevante opțiuni. Această capacitate este deosebit de valoroasă în sectoare precum sănătate, unde recomandările personalizate pot îmbunătăți semnificativ rezultatele pacienților, sugerând tratamente sau măsuri preventive adaptate datelor individuale.

Inteligența artificială și învățarea automată permit sistemelor nu doar să învețe din seturi vaste de date, ci și să se adapteze în timp real la comportamentele și preferințele utilizatorilor în schimbare. Această adaptabilitate este crucială pentru menținerea angajamentului și satisfacției utilizatorilor într-o lume în care preferințele se pot schimba rapid și imprevizibil.

În plus, implicațiile etice și sociale ale Sistemelor de Recomandare necesită o examinare atentă. Pe măsură ce aceste sisteme exercită o influență semnificativă asupra a ceea ce utilizatorii văd, cumpără sau consumă, există o nevoie imperativă de a asigura că recomandările sunt corecte, nepartinitoare și transparente. Abordarea problemelor precum biasul algoritmic, confidențialitatea datelor și încrederea utilizatorilor reprezintă o parte esențială a dezvoltării și rafinării continue a cercetării în domeniul

Sistemelor de Recomandare.

În cercetarea academică, Sistemele de Recomandare oferă un teren fertil pentru explorarea abordărilor interdisciplinare care integrează perspective din informatică, psihologie, economie și sociologie. Înțelegerea nuanțelor comportamentului utilizatorilor și a formării preferințelor necesită o perspectivă holistică care ia în considerare factorii cognitivi și sociali, îmbogățind astfel dimensiunile teoretice și practice ale domeniului.

În cele din urmă, motivația de a aprofunda domeniul Sistemelor de Recomandare este înrădăcinată în potențialul lor profund de a transforma modul în care interacționăm cu informațiile și luăm decizii. Extinderea limitelor a ceea ce pot realiza aceste sisteme contribuie la un viitor digital mai personalizat, eficient și centrat pe utilizator.

1.2 Contribuții originale

Principalele contribuții originale ale acestei teze sunt prezentate astfel:

- Prima abordare [6], introdusă în Capitolul 3, prezintă o comparație detaliată a diverselor metrici de similaritate utilizate în filtrarea colaborativă bazată pe memorie (MBCF), oferind perspective nuanțate asupra performanței acestora în diverse contexte și luând in considerare diferite caracteristici ale seturilor de date. Această analiză comparativă nu doar identifică măsurile de similaritate optime, ci oferă și recomandări practice pentru proiectanții sistemelor.

- Recunoscând limitările evaluărilor numerice ale utilizatorilor în a exprima opinii nuanțate despre produse, utilizarea descrierilor textuale devine esențială pentru îmbunătățirea procesului de recomandare. În consecință, abordarea [8] analizează sistematic informațiile textuale, calculând scoruri de sentiment pentru a interpreta mai bine opiniile utilizatorilor. Contribuția originală constă în introducerea unei măsuri de similitudine bazată pe sentiment, Atracție-Relevanță-Popularitate (ARP). Spre deosebire de măsurile convenționale de similitudine, ARP utilizează evaluări bazate pe sentiment pentru a determina similaritatea între utilizatori, oferind o perspectivă unică care îmbunătățește semnificativ performanța SR. Este important de menționat că Modulul de Evaluare a Sentimentelor a fost parte din cercetarea comună [8] și nu este o contribuție originală a acestei teze.

- Un cadru de validare este propus în [15] pentru a evalua noua măsură de similaritate ARP și pentru a demonstra că poate fi utilizată cu succes în locul celor tradiționale. Majoritatea abordărilor existente în literatură validează noile măsuri doar în termeni de tehnici de evaluare, de exemplu: precizie, acuratețe sau eroare absolută medie. Contribuția originală din cadrul abordării [15] constă în propunerea unui cadru de validare pentru noua măsură de similaritate ARP, incluzând patru din cele cinci componente esențiale ale procesului de validare: verificarea axiomelor metricii, utilitatea, expresivitatea și corelația cu alte măsuri. A cincea componentă de validare, robustetea la zgomot, este doar sumarizată în această teză deoarece nu este o contribuție originală. Detaliile privind experimentele numerice și rezultatele obținute pot fi consultate în cercetarea comună [15].

- Secțiunea 5.1 din Capitolul 5 prezintă o abordare [19] care are ca scop optimizarea tehnicii de recomandare KNN prin utilizarea analizei sentimentelor (SA) în faza de preprocesare a datelor

din procesul de recomandare. SA este aplicată pentru clasificarea recenziilor textuale ale restaurantelor în pozitive și negative. Setul de date rezultat este transmis unui sistem de recomandare care, utilizând algoritmul KNN, prezice evaluarea pentru un restaurant nevizitat și generează o listă de restaurante recomandate pentru utilizator. Această abordare a depășit rezultatele obținute când pasul SA nu a fost considerat în procesul de recomandare.

- Abordarea [7], prezentată în Secțiunea 5.2 din Capitolul 5, pledează pentru implementarea unei tehnici de filtrare colaborativă KNN bazată pe lexicon, integrând evaluările sentimentului calculate în etapa de determinare a vecinătății. Această metodologie prezintă două contribuții originale principale. În primul rând, se realizează o analiză detaliată a informațiilor textuale asociate fiecărui articol și apoi se aplică Lexiconul de Analiză a Sentimentelor Vader [12] pentru a deriva evaluările sentimentului. În al doilea rând, setul de date bazat pe lexicon este transmis tehnicii de filtrare colaborativă KNN bazată pe utilizatori. Experimentele realizate arată că sistemul de recomandare bazat pe text obține recomandări precise pentru utilizatori.

- În Secțiunea 5.3 din Capitolul 5, este introdus un sistem de recomandare nou, destinat cititorilor New York Times, utilizând o perspectivă de învățare automată nesupervizată [18]. În primul rând, abordarea propusă integrează o perspectivă nesupervizată prin incorporarea algoritmului K-Means în tehnica tradițională de recomandare KNN. Această fuziune îmbunătățește capacitatea sistemului de a identifica clustere de articole bazate pe subiecte sau teme similare, rafinând astfel procesul de recomandare. Mai mult, sistemul de recomandare este proiectat și personalizat meticulos pentru a se potrivi cu gama extinsă de articole disponibile pe portalul New York Times. Adaptând sistemul la conținutul și publicul specific al New York Times, eficacitatea și relevanța recomandărilor sunt substanțial crescute, asigurând o experiență mai îmbogățită pentru utilizatori. În plus, originalitatea lucrării se extinde la experimentele numerice riguroase realizate pe multiple seturi de date provenite din arhivele New York Times, cuprinzând articole din diverse perioade. Aceste experimente servesc la validarea eficienței și robusteții atât a proceselor de clusterizare, cât și a celor de recomandare, oferind dovezi empirice ale performanței sistemului în diverse seturi de date și contexte temporale.

Capitolul 2

Concepte Cheie și Algoritmi în Sistemele de Recomandare

Sistemele de recomandare (RS) cuprind instrumente software și metodologii concepute pentru a oferi utilizatorilor îndrumare în diverse scenarii de luare a deciziilor: ce articole să cumpere, ce cărți merită citite sau la ce restaurant să ia masa. RS și-au găsit inspirația în binecunoscutul comportament de turmă: oamenii au adesea încredere în recomandările altora atunci când iau decizii în viaţa de de zi cu zi.

În ultimii ani, RS s-au dovedit a fi o soluție eficientă pentru problema supraîncărcării de informații. Practic, un RS conduce utilizatorul către elemente noi, încă nedescoperite, care cel mai probabil sunt de interes pentru nevoile curente ale utilizatorului. Elementul recomandat utilizatorilor de către RS face parte cel mai adesea dintr-o categorie specifică și singulară (de exemplu, filme, muzică, știri sau restaurante). Sugestiile personalizate sunt prezentate sub formă de matrice ordonată de elemente. La calcularea clasamentului, RS utilizează preferințele utilizatorilor, cum ar fi evaluările acordate produselor sau chiar navigarea către o anumită pagină de produs. După ce primește recomandările, utilizatorul le poate parcurge și poate decide dacă le acceptă sau le refuză. Ulterior, utilizatorul poate oferi un feedback implicit sau explicit, fie imediat, fie în timpul unei interacțiuni ulterioare. Aceste interacțiuni cu utilizatorul și feedback-ul acestuia pot fi reținute și utilizate pentru a produce noi recomandări în timpul unor interacțiuni viitoare între utilizator și sistem.

În acest capitol, este prezentat un sumar teoretic al diferitelor tehnici de recomandare şi sunt identificate punctele forte și punctele slabe ale acestor algoritmi de recomandare.

Un RS colectează în mod activ și continuu diverse tipuri de informații cu scopul de a oferi sugestii. Aceste informații se referă atât la produsele care urmează să fie recomandate, cât și la utilizatorii care primesc sugestiile. În general, informațiile utilizate de RS constau în: articole, utilizatori și relațiile dintre articole și utilizatori.

Articolele recomandate de RS posedă caracteristici distincte, cum ar fi complexitatea, valoarea sau utilitatea. Valoarea unui articol poate fi fie pozitivă, indicând relevanța acestuia pentru utilizator, fie negativă, semnificând o selecție necorespunzătoare din partea utilizatorului. Unele elemente pot prezenta o complexitate și o valoare scăzute, cum ar fi articolele de știri, site-urile web și filmele, în timp ce alte elemente au o complexitate și o valoare mai mari: aparate foto, calculatoare. Elementele considerate cele mai importante sunt: polițele de asigurare, investițiile financiare, călătoriile planificate și locurile de muncă [20].

Utilizatorii pot avea obiective și caracteristici diferite. Pentru a personaliza sugestiile și pentru a îmbunătăți experiența utilizatorilor, RS exploatează informațiile colectate despre utilizator. Alegerea informațiilor care trebuie modelate variază în funcție de tehnica de recomandare utilizată. De exemplu, în filtrarea colaborativă, profilu utilizatorului cuprinde o listă care conține evaluările utilizatorilor cu privire la diverse articole. În schimb, într-un RS demografic, atributele socio-demografice precum vârsta, sexul, educația și experiența profesională sunt încorporate în profilurile utilizatorilor [20].

Preferințele și cerințele utilizatorului sunt stocate în modelul lor. Un sistem de recomandare acționează ca un instrument care formulează sugestii prin crearea și utilizarea acestor modele. Având în vedere că personalizarea depinde de existența unui model de utilizator eficient, acesta rămâne esențial în procesul de recomandare. De exemplu, într-o abordare care foloseşte filtrarea colaborativă, utilizatorul este caracterizat fie direct prin evaluările pe care le acordă articolelor, fie sistemul deduce un vector de valori ale factorilor din aceste evaluări, reflectând variațiile în modul în care utilizatorii cântăresc fiecare factor în modelul lor [20].

Utilizatorii pot fi caracterizați pe baza modelelor lor comportamentale, cum ar fi modelele de navigare pe site sau modelele de căutare a călătoriilor. Baza de informații despre utilizatori poate include, de asemenea, relații între aceștia. Prin urmare, RS va utiliza aceste date pentru a sugera elemente similare preferate de alți utilizatori de încredere [20].

Relațiile utilizator-element reprezintă înregistrări de tip jurnal care conțin detalii semnificative generate în timpul interacțiunilor utilizator-sistem, servind ca date de intrare pentru algoritmul de recomandare. În mod obișnuit, evaluările constituie cel mai frecvent tip de date colectate de sistem [20]. Ricci et al. [20] oferă o clasificare detaliată a ratingurilor: numerice, ordinale, binare, respectiv unare.

Bobadilla et al. [3] propune următorii pași de urmat în dezvoltarea unui RS:

- Structura setului de date (de exemplu, evaluări numerice sau textuale).
- Tehnica de filtrare selectată: colaborativă, bazată pe conținut, hibridă, learning to rank, context-aware, social-based.
- Modelul selectat: bazat pe memorie sau pe model.
- Sparsitatea setului de date.
- Scalabilitatea dorită.
- Performanța RS: timpul de execuţie și utilizarea memoriei.

Capitolul 3

Rolul Măsurilor de Similaritate în Optimizarea Acurateței Recomandărilor

Alegerea măsurilor de similaritate influențează în mod semnificativ eficacitatea și calitatea recomandărilor generate de sistemele de recomandare.

Acest capitol își propune să cerceteze rolului fundamental al măsurilor de similaritate în îmbunătățirea performanțelor sistemelor de recomandare.

Unul dintre obiective este de a investiga influența diferitelor măsuri de similaritate utilizate pe scară largă, inclusiv, dar fără a se limita la similaritatea cosinusului, coeficientul de corelație Pearson și indicele Jaccard, asupra procesului de recomandare. O analiză riguroasă a acestor metrici a identificat punctele forte, punctele slabe și aplicabilitatea lor în diferite scenarii de recomandare.

În plus, acest capitol va examina modul în care selecția și integrarea adecvată a măsurilor de similaritate pot duce la optimizarea acurateței recomandărilor.

Pentru a atinge aceste obiective, s-a efectuat o revizuire și o analiză amănunțită a studiilor relevante care au investigat impactul măsurilor de similaritate asupra acurateței recomandărilor, inclusiv o evaluare a metodologiilor, a rezultatelor și a limitărilor ale acestora. De asemenea, acest capitol prezintă o analiză comparativă a performanțelor diferitelor măsuri de similaritate aplicate pe algoritmi de filtrare colaborativă bazată pe memorie: scenarii bazate pe utilizator (UBCF), respectiv bazate pe elemente/articole (IBCF).

Au fost efectuate mai multe experimente numerice, luând în considerare următoarele măsuri de similaritate: PIP [2], coeficientul de corelație Pearson (PCC), coeficientul de corelație Pearson constrâns (CPC), similitudinea Cosinus (COS), Cosinus ajustat (ACOS), indicele Jaccard (JAC), distanța euclidiană (EUC) și coeficientul de rang Spearman (SRC) [**?**] și tehnici de filtrare colaborativă bazate pe memorie (UBCF și IBCF).

În etapa de evaluare, au fost utilizate ca măsuri de evaluare precizia și eroarea medie absolută.

Pentru experimentele numerice au fost alese două seturi de date, care sunt diferite din punct de vedere al dimensiunii și al sparsităţii.

Setul de date MovieLens 1M[1] conține 1 milion de recenzii aplicate pe 4 000 de filme de către 6

[1] https://grouplens.org/datasets/movielens/

000 de utilizatori. Setul de date DataFiniti Hotel Reviews[2] conține 10 000 de recenzii pentru 1 670 de hoteluri. Ambele seturi de date iau în considerare recenzii ale utilizatorilor cu valori de la unul la cinci.

Din punct de vedere al sparsităţii, MovieLens are 95,83%, în timp ce DataFiniti Hotel Reviews 99,91%.

Rezultate și Observații

În cadrul studiului prezentat, au fost efectuate mai multe experimente pentru a oferi un răspuns la un set de întrebări esențiale în abordările de filtrare colaborativă bazate pe memorie. Principala dificultate în proiectarea unui sistem de recomandare constă în alegerea corectă a măsurii de similaritate.

Tabela 3.1: Constatări privind măsurile de similaritate.

Caracteristică	**Comentarii**
Set de date	MovieLens și DataFiniti Hotel Reviews
Impactul sparsităţii si al dimensiunii asupra IBCF	JAC are cele mai bune rezultate pentru seturi de date mici cu sparsitate mare, deoarece compară prezența sau absența ratingurilor (valori binare), ceea ce îl face aplicabil atunci când se utilizează seturi de date cu o proporție mare de elemente neevaluate. SRC se potrivește abordării bazate pe itemi pentru seturi mari de date.
Impactul sparsităţii si al dimensiunii asupra asupra UBCF	PIP se potrivește pentru seturi de date mari și cu sparsitate scăzută. Mai multe similitudini pot fi utilizate pentru seturi de date mici și cu sparsitate ridicată (COS, SRC).

Seturile de date utilizate, MovieLens[3] și DataFiniti - Hotel Reviews[4], au fost alese din punct de vedere al dimensiunii și al caracteristicilor de sparsitate diferite. Rezultatele experimentelor numerice efectuate au condus la următoarele concluzii. În ceea ce privește seturile mari de date și o mai mică sparsitate a datelor, similitudinea PIP se potrivește contextului bazat pe utilizator, în timp ce coeficientul de rang al lui Spearman (SRC) ar putea fi o selecție adecvată pentru contextul bazat pe elemente. În schimb, în cazul în care există un set de date mai mic cu o sparsitate mare, similaritatea Jaccard se potrivește contextului bazat pe elemente. Pentru scenariul bazat pe utilizator, pot fi alese mai multe similarităţi (COS, SRC), în funcție de dimensiunea vecinătății. În plus, în această analiză au fost discutate principalele caracteristici ale măsurilor de similaritate care influențează pozitiv procesul

[2]https://data.world/datafiniti/hotel-reviews
[3]https://grouplens.org/datasets/movielens/
[4]https://data.world/datafiniti/hotel-reviews

de recomandare.

Tabelul 3.1 reflectă rezumatul analizei prezentate, evidențiind setul de date utilizat, algoritmii și concluzia privind impactul similarității în ceea ce privește dimensiunea și sparsitatea datelor.

În concluzie, acest capitol are ca scop evidenţierea rolul critic al măsurilor de similaritate în optimizarea acurateței recomandărilor. Prin explorarea diverselor măsuri de similaritate, a influenței lor asupra rezultatelor recomandării și a strategiilor pentru utilizarea lor eficientă, ne propunem să contribuim la progresul sistemelor de recomandare și să oferim informații valoroase cercetătorilor, practicienilor și dezvoltatorilor care activează în acest domeniu.

Pentru a îmbunătăți în continuare analiza măsurilor de similaritate, o direcție ar fi explorarea unor parametri de evaluare suplimentari, cum ar fi: acoperirea - pentru a măsura proporția de elemente pentru care se pot face recomandări; serendipitatea - pentru a evalua cât de surprinzătoare și inedite sunt recomandările pentru utilizator; diversitatea - pentru a evalua varietatea recomandărilor pentru a evita redundanța; măsurile online - pentru a analiza parametrii de performanță în timp real, cum ar fi click-urile sau parametrii de interacțiune cu utilizatorul. Aceste măsurători pot oferi o evaluare mai cuprinzătoare a eficacității sistemului de recomandare, dincolo de precizie și MAE, asigurând o abordare echilibrată și centrată pe utilizator.

Capitolul 4

Măsura de Similaritate Atractivitate-Relevanţă-Popularitate (ARP)

Sistemele de recomandare sunt instrumente care interpretează preferințele utilizatorilor în încercarea de a genera sugestii adecvate. Studiile de cercetare tind să concluzioneze că evaluările numerice ale utilizatorilor nu sunt suficient de puternice pentru a exprima cu adevărat preferințele utilizatorilor. Pe de altă parte, recenziile bazate pe text pot exprima caracteristici precum sentimente, opinii sau atitudini, care sunt mai promițătoare pentru extragerea de informații valoroase. O recenzie bazată pe text poate fi utilizată pentru a defini sentimentul sau opinia generală cu privire la un articol. Prin urmare, extragerea sentimentului sau a polarității informațiilor textuale se dovedește a fi o componentă esențială a unui sistem de recomandare.

Măsura de similaritate este un concept-cheie într-o gamă largă de procese din domenii, cum ar fi prelucrarea limbajului natural, clustering sau sistemele de recomandare. În ultimii ani, au fost concepute noi măsuri de similaritate în diferite contexte. În prezent, există o lipsă profundă în ceea ce priveşte etapele de validare și evaluare a noilor similarităţi. În general, noile măsuri sunt validate prin experimente numerice care utilizează diferite seturi de date și în termeni de măsuri de evaluare, cum ar fi: acuratețea, precizia sau eroarea medie absolută. Dar acest lucru nu este suficient, aşadar este necesar un proces de validare mai complex.

Acest capitol prezintă proiectarea unei noi măsuri de similaritate, **Atractivitate-Relevanţă-Popularitate (ARP)**, care a rezultat din integrarea tehnicilor de analiză a sentimentelor (SA) în procesul de recomandare pentru a crește precizia elementelor sugerate [8]. În plus, este introdus un cadru de validare pentru a evalua valoarea adăugată a ARP și pentru a dovedi că poate fi utilizată cu încredere în locul măsurilor tradiționale de similaritate [15]. Procesul de validare constă în cinci etape principale: verificarea condițiilor unei metrici, utilitatea, expresivitatea, corelațiile cu alte măsuri și robustețea zgomotului. Au fost luate în considerare și analizate mai multe criterii, cum ar fi: măsura poate fi aplicată pe seturi de date cu diferite tipuri de date (de exemplu, caracteristici numerice, categorice) sau cât de eficientă este similaritatea nouă?

Prezentarea Metodologiei

În Figura 4.1 se pot distinge principalele componente ale sistemului proiectat.

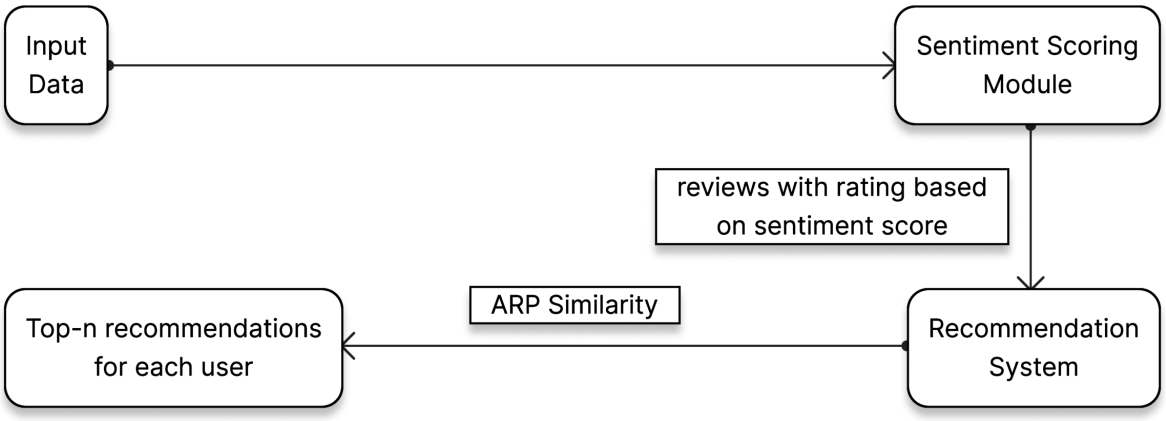

Figura 4.1: ARP: Metodologia

În primul rând, datele colectate sunt transmise către modulul Sentiment Score. Pentru fiecare recenzie bazată pe text, se calculează un scor de sentiment pe baza următoarelor etape. Într-o primă etapă, scorul de polaritate al unui cuvânt este obţinut cu ajutorul softului SentiWordNet[1], rezultând un scor pozitiv sau negativ.

Ulterior, scorul de sentiment al unei recenzii este calculat prin însumarea scorurilor de sentiment ale tuturor cuvintelor care o compun. Se aplică apoi o funcție de evaluare pentru a converti scorul de sentiment într-o evaluare reală cu valori de la unu la cinci pentru fiecare recenzie bazată pe text. Detaliile modulului Sentiment Score sunt descrise în abordarea comună de cercetare [**?**]. Este important de remarcat faptul că acest aspect face parte dintr-o abordare de cercetare comună [8] și nu reprezintă o contribuție originală a acestei teze.

În plus, ratingurile de sentiment atribuite fiecărei recenzii înlocuiesc ratingul original al utilizatorului și sunt transmise componentei sistemului de recomandare. Algoritmul de filtrare colaborativă KNN bazat pe utilizator este selectat ca tehnică de recomandare. Similaritatea dintre doi utilizatori este calculată cu ajutorul modelului de similaritate bazat pe sentiment ARP. În cele din urmă, componenta Sistem de Recomandare produce o listă de recomandări pentru fiecare utilizator.

ARP: Principii de proiectare

Măsura ARP bazată pe sentiment este definită pe baza a trei factori de similaritate: Atractivitatea, Relevanța și Popularitatea.

Atractivitatea unei recenzii depinde de numărul de scoruri pozitive și negative ale cuvintelor care o compun. În acest context, ea marchează cât de atractiv ar putea fi un articol pentru utilizatori.

Conceptul de popularitate a fost definit din trei perspective: pentru o recenzie, pentru un utilizator și pentru un utilizator în raport cu un alt utilizator. Factorul de popularitate relevă cât de mult se abate recenzia/utilizatorul de la medie.

[1] http://ontotext.fbk.eu/sentiwn.html

Factorul de relevanță se bazează pe atractivitate și poate fi definit pentru un utilizator, o recenzie și un utilizator în raport cu un alt utilizator. Relevanța indică abaterea medie a recenziei j dată de utilizatorul i în ceea ce privește atractivitatea.

Măsura de similaritate ARP propusă are valori cuprinse între [-1,1], unde o valoare apropiată de 1 indică o mai mare similaritate între utilizatori. În comparație cu măsurile tradiționale de similaritate care utilizează recenzia numerică cu valori de la unu la cinci a utilizatorilor, similaritatea ARP este formulată exclusiv sentimentul calculat anterior pentru recenziile bazate pe text.

ARP: Aplicație în Sistemele de Recomandare

În procesul de recomandare, se utilizează algoritmul KNN CF bazat pe utilizatori. O etapă importantă în dezvoltarea unui RS care utilizează tehnica de filtrare colaborativă este selectarea măsurii de similaritate corespunzătoare. Pentru a determina grupul de k vecini ai utilizatorului țintă, măsura ARP calculează similarităţile dintre recenziile bazate pe sentiment ale utilizatorilor. Pe baza recenziei prezise, se generează o listă de recomandare top-n pentru utilizatorul țintă.

În experimentele numerice au fost utilizate seturile de date Yelp Restaurants Reviews[2] și Datafiniti Hotel Reviews[3]. În ambele cazuri, s-au utilizat recenziile bazate pe sentiment. La fel ca în cazul abordării recenziilor bazate pe sentiment, pentru faza de testare a sistemului de recomandare au fost utilizate 20% din date.

Primul experiment numeric are ca scop determinarea performanței sistemului de recomandare folosind măsura de similaritate ARP. Experimentele sunt efectuate pe ambele seturi de date și au fost luate în considerare diferite valori pentru numărul de vecini k și numărul de recomandări n. Rezultatele pentru setul de date Yelp, arată că cele mai bune valori, în ceea ce privește MAE și RMSE, au fost obținute pentru $k = 5$ și cele mai bune 15 recomandări (MAE=0,03; RMSE=0,18), respectiv pentru $k = 10$ și cele mai bune 5 recomandări (MAE=0,09; RMSE=0,10).

Rezultatele pentru setul de date DataFiniti arată că cele mai bune valori, în termeni de MAE, au fost obținute pentru $k = 5$ și primele 15 recomandări (MAE=0,07), respectiv pentru $k = 10$ și primele 5 recomandări (MAE=0,17). Cele mai bune rezultate în ceea ce privește RMSE au fost înregistrate pentru $k = 5$ și primele 10 recomandări (RMSE=0,30), respectiv pentru $k = 10$ și primele 5 recomandări (RMSE=0,37).

Al doilea experiment are ca scop compararea performanțelor ARP cu mai multe măsuri tradiționale de similaritate: PIP [2], coeficientul de corelație Pearson (PCC), similitudinea cosinus (COS), indicele Jaccard (JAC) și coeficientul de rang Spearman (SRC) [**?**]. Având în vedere cele mai bune rezultate obținute în primul experiment, cel de-al doilea este realizat pe $k = 5$ vecini și primele 15 recomandări, respectiv $k = 10$ și primele 5 recomandări. Pentru setul de date Yelp, rezultatele arată că măsura de similaritate SRC oferă cele mai bune valori (MAE=0,01 și RMSE=0,12). Măsura ARP produce, de asemenea, valori comparabile și favorabile, cu MAE=0,03 și RMSE=0,18, ceea ce o situează pe locul al doilea în topul celor mai bune măsuri.

Pentru setul de date DataFiniti, luând în considerare configurația k=5 și n=15, măsura de similaritate SRC obține cele mai bune valori, cu MAE=0,02 și RMSE=0,13. Măsura ARP oferă, de asemenea, rezultate bune, cu MAE=0,03 și RMSE=0,18, apropiindu-se foarte mult de performanța măsurii de similaritate SRC.

[2]https://www.yelp.com/dataset
[3]https://data.world/datafiniti/hotel-reviews

Criterii de Validare pentru Similaritatea ARP

Pentru a dovedi că măsura de similaritate ARP poate fi utilizată cu încredere ca o alternativă la măsurile de similaritate populare în contextul filtrării colaborative, se propune un cadru de validare. Procesul de validare este descris în figura 4.2.

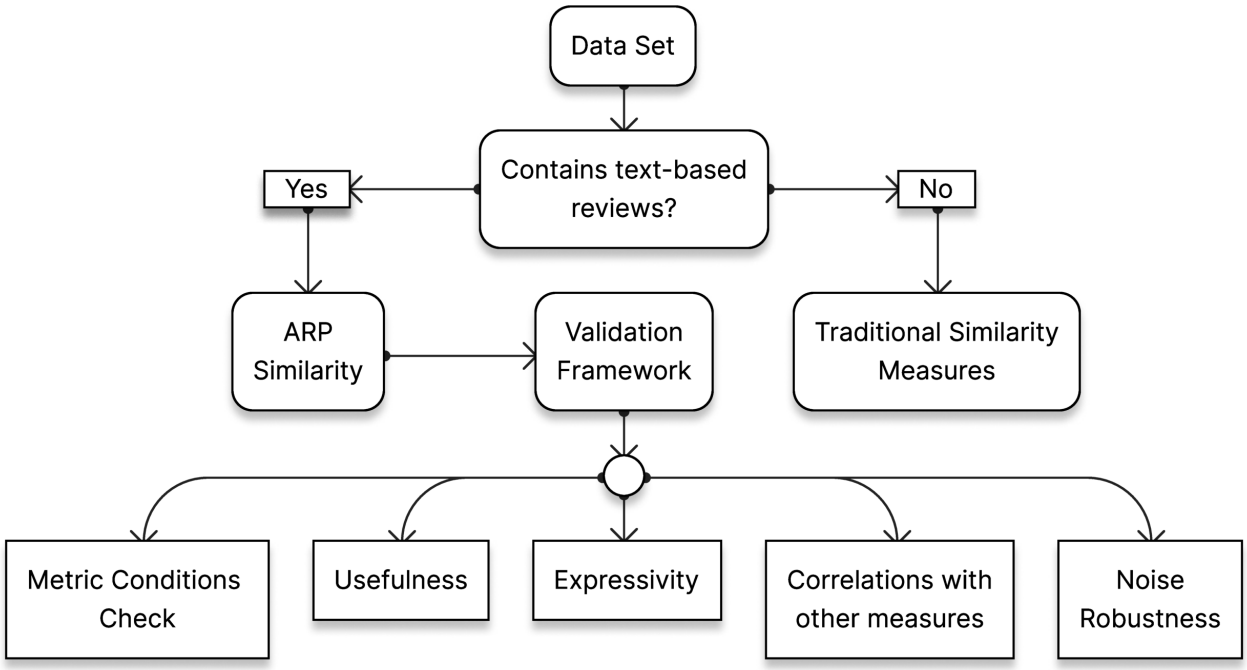

Figura 4.2: ARP: Criterii de Validare

În faza de colectare a datelor, trebuie să se verifice dacă setul de date conține recenzii bazate pe text pentru articole. Dacă este cazul, atunci măsura ARP poate fi utilizată pentru a îmbunătăți predicția ratingului unui sistem de recomandare și se poate aplica cadrul de validare. În caz contrar, în procesul de recomandare trebuie utilizate alte măsuri de similaritate bine cunoscute, cum ar fi coeficientul de corelație Pearson sau similaritatea cosinusului.

Cadrul de validare este format din următoarele componente: Verificarea Condițiilor unei metrici, Verificarea utilității, Expresivitatea, Corelațiile cu alte măsuri de similaritate și Robustețea la zgomot.

Concluzii

Acest capitol prezintă o nouă măsură de calcul a similarităţii utilizatorilor bazată pe sentiment (ARP), care exploatează opiniile utilizatorului derivate din recenziile sale bazate pe text. Măsura ARP a fost testată pentru algoritmul de filtrare colaborativă K Nearest Neighbors, utilizând două seturi de date publice: Yelp Restaurants Reviews și Datafiniti Hotel Reviews. Performanța ARP a fost comparată cu măsurile tradiționale de similaritate în contextul CF. Rezultatele arată că ARP poate înlocui cu succes măsurile tradiționale de similaritate, cum ar fi coeficientul de corelație Pearson, cosinus sau PIP [2].

Chiar dacă nu sunt posibile comparații cu lucrări conexe din literatura de specialitate [9, 5, 17, 11, 13, 14], avantajul clar al modelului propus este că algoritmul de recomandare nu suferă nicio

adaptare, astfel încât măsura ARP poate fi integrată cu ușurință în procesul existent. Un aspect foarte important este faptul că similaritatea ARP poate fi utilizată și pentru date de intrare care nu conțin evaluări numerice, ci doar descrieri textuale ale elementelor. Acest aspect nu este valabil pentru măsurile tradiționale.

În plus, a fost propus un cadru de validare pentru ARP, luând în considerare cinci componente independente: utilitatea, expresivitatea, corelația cu alte măsuri, verificarea condițiilor unei metrici și robustețea zgomotului.

Planurile de extindere includ aplicarea similitudinii ARP pe mai multe seturi de date pentru a studia efectele dimensiunii și ale caracteristicilor de sparsitate asupra performanței sale. De asemenea, pot fi explorate și alte caracteristici relevante ale informațiilor textuale pentru un element.

Capitolul 5

Cadre de Optimizare pentru Tehnicile de Recomandare

5.1 Valorificarea Analizei Sentimentelor pentru Îmbunătățirea Predicțiilor de Rating în Sistemele de Recomandare

Scopul acestei secțiuni este de a introduce o perspectivă originală [19] care îmbunătățește rezultatele CF prin încorporarea unei etape de preprocesare a datelor bazată pe analiza sentimentelor în procesul de recomandare. Această metodă a fost testată prin scenarii experimentale, după cum urmează: clasificatorul de sentiment efectuează primul nivel de filtrare, în timp ce algoritmul CF este utilizat la al doilea nivel de filtrare. Rezultatele abordării propuse sunt comparate cu cele produse de algoritmul CF de bază, în care nu se iau în considerare rezultatele sentimentului. Rezultatul final este o listă mai precisă de recomandări generată pentru utilizatori.

Abordarea propusă integrează tehnicile SA într-un algoritm CF. În figura 5.1, sunt descrise principalele componente și interacțiuni ale sistemului propus. Pe baza setului de date de intrare, clasificatorul de sentiment produce ca ieșire un set de date etichetate pozitiv/negativ. Acesta din urmă va fi transmis sistemului de recomandare pentru a genera o listă de recomandări.

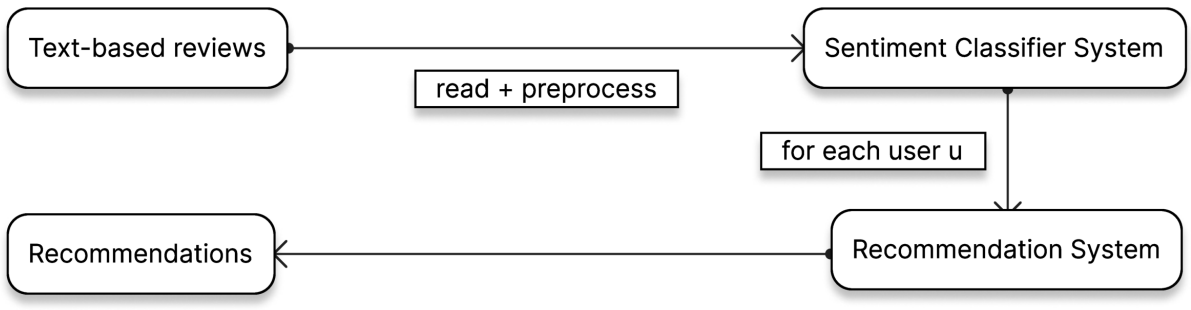

Figura 5.1: Sistemul de Recomandare bazat pe Sentiment

Rezultate

Setul de date de intrare este format din 8.539 de recenzii etichetate pozitiv sau negativ, păstrând următoarele caracteristici de interes: ID-ul afacerii, stelele acordate de un utilizator (valori de la unu la cinci), ID-ul utilizatorului și recenzia în format text.

În faza de evaluare, 20% din setul de date de intrare a fost utilizat ca set de testare.

Pentru a evalua sistemul de recomandare a restaurantelor bazat pe sentiment (SA-enhanced RRS), au fost efectuate mai multe experimente, luând în considerare diferite dimensiuni ale vecinătăților ($k = 5$ și $k = 10$) și diferite numere de recomandări generate ($n = 3$, $n = 5$, $n = 10$ și $n = 15$). În acest context, rezultatele obținute au fost comparate în continuare cu scenariul trivial fără a lua în considerare noile etichete de sentiment (RRS de bază). Rezultatele arată că etapa de analiză a sentimentelor crește calitatea recomandărilor, deoarece s-au obținut valori mai bune pentru măsurile de evaluare precizie, recall, F-Score și MAE.

În faza de recomandare, numai recenziile etichetate cu un sentiment pozitiv, în etapa de analiză a sentimentelor, sunt considerate relevante și sunt utilizate în calculul similarității. Un avantaj este faptul că componenta SA e implementată independent de procesul de recomandare, iar datele obţinute sunt utilizate ca date de intrare pentru sistemul de recomandare.

5.2 O Abordare de Filtrare Colaborativă bazată pe Lexicon pentru Sistemele de Recomandare

Această secțiune prezintă o abordare originală [7] concepută pentru a capta interesele utilizatorilor din recenziile articolelor bazate pe text pentru a produce predicții bune de rating pentru articole și recomandări generate cu precizie pentru utilizatori. Descrierile articolelor sunt transmise unui lexicon SA, care produce un scor de sentiment care indică polaritatea textului (pozitiv, negativ sau neutru). Pe baza scorului de sentiment, a fost aplicat un algoritm de CF bazat pe utilizator KNN. RS utilizează exclusiv scorurile de sentiment (denumite ratinguri de sentiment), în loc de ratingurile numerice. Rezultatele au dovedit un impact pozitiv al abordării bazate pe text asupra performanței sistemului de recomandare.

Deoarece descrierile elementelor bazate pe text dezvăluie informații mai valoroase în comparație cu simplele evaluări numerice, abordarea propusă se concentrează pe utilizarea exclusivă a informațiilor textuale în crearea sistemului de recomandare. Datele textuale sunt exploatate cu ajutorul unei tehnici bazate pe lexicon pentru a determina scorul de polaritate al unei recenzii. Scorurile rezultate sunt evaluările de sentiment luate în considerare pentru algoritmul de filtrare colaborativă kNN bazat pe utilizator.

Figura 5.2 prezintă arhitectura propusă pentru sistemul proiectat. După faza de colectare a datelor, recenziile articolelor bazate pe text servesc drept date de intrare pentru un lexicon de sentiment care determină o evaluare a sentimentului pentru un articol. Setul de date îmbogățit cu ratingul de sentiment calculat este transmis ulterior unui sistem de recomandare.

Abordarea propusă utilizează, pentru sarcina de analiză a sentimentului, un lexicon de sentiment, care a fost selectat pe baza comparației complexe și amănunțite prezentate în [12]. Lexiconul Vader Sentiment a fost comparat cu mai multe din literatura de specialitate (Linguistic Inquiry Word Count, General Inquirer, Affective Norms for English Words, SentiWordNet, SenticNet, Word-Sense

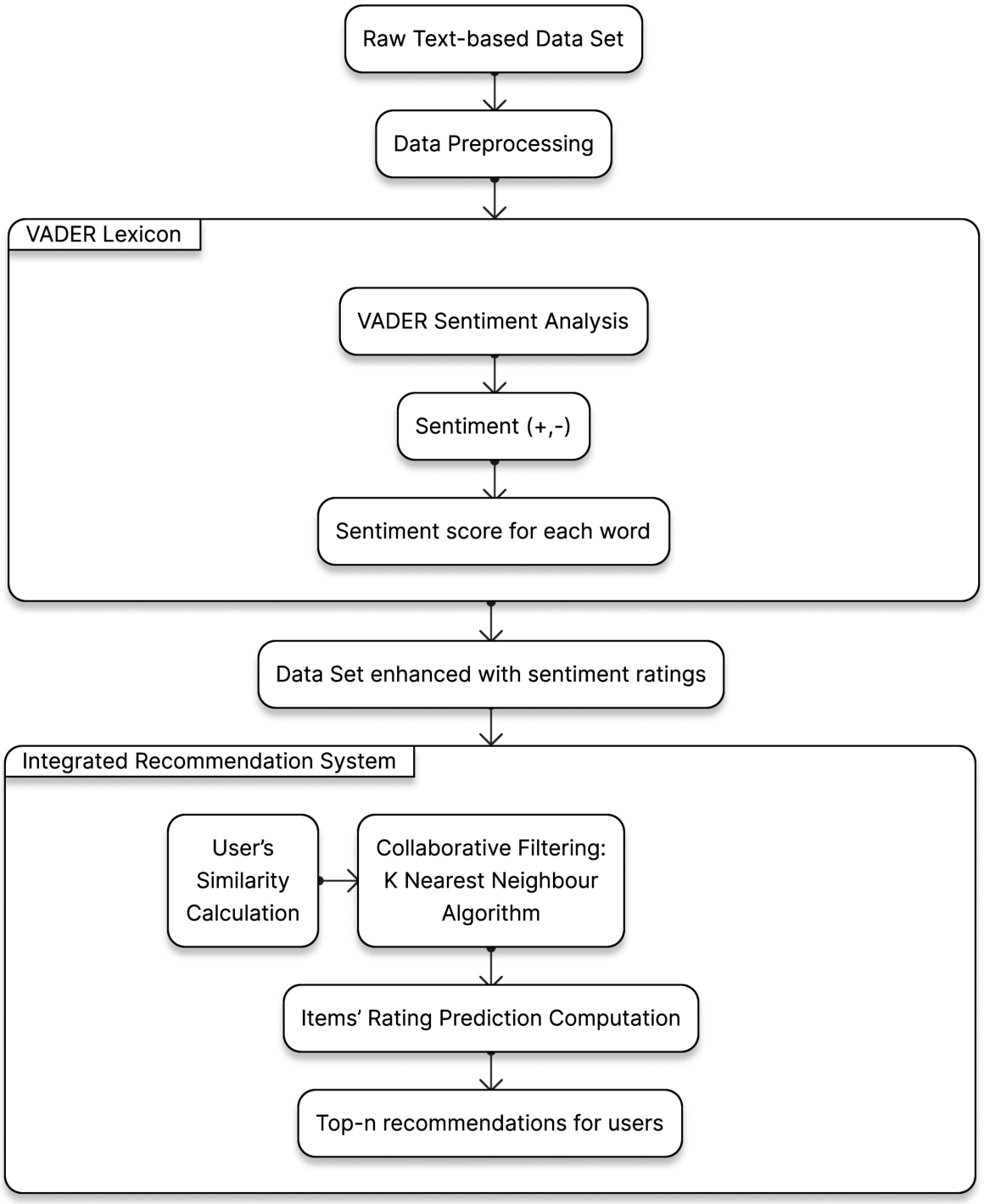

Figura 5.2: Arhitectura Sistemului de Recomandare bazat pe Lexicon.

Disambiguation) și a produs, în majoritatea cazurilor, cele mai bune rezultate.

Setul de date care conține în plus recenziile cu scoruri de sentiment reprezintă datele de intrare pentru sistemul de recomandare. Algoritmul clasic KNN CF este apoi aplicat ca tehnică de recomandare [19].

Experimente Numerice

Pentru a evidenția valoarea abordării propuse în îmbunătățirea acurateței de predicție a ratingului, au fost efectuate mai multe experimente numerice pe trei seturi de date care conțin recenzii bazate pe text pentru articole.

Pentru determinarea vecinătății în KNN, au fost aplicate diferite măsuri de similaritate populare din literatura de specialitate: Pearson Correlation Coefficient (PCC), Cosine (COS), Euclidian (EUC), Constrained Pearson Coefficient (CPC), Spearman Rank Coefficient (SRC), Jaccard Similarity (JAC) [1], [21] și PIP [2]. Sunt concepute scenarii independente pentru diferite valori ale lui k (dimensiunea vecinătății) și n (numărul de recomandări generate).

În procesul de evaluare, se calculează măsurile MAE și RMSE pentru a stabili acuratețea recomandărilor generate.

În plus, abordarea propusă bazată pe lexicon este comparată cu o altă abordare KNN CF bazată pe text, descrisă în [22], în ceea ce privește măsura de performanță RMSE. Ambele abordări utilizează recenzii bazate pe text în loc de recenzii numerice, iar experimentele sunt efectuate pe setul de date Rotten Tomato Critic Reviews. Din păcate, în cazul abordării din [22], detaliile privind valorile alese pentru dimensiunea vecinătății (k) și numărul de recomandări (n) nu sunt comunicate în configurația experimentală. Deși ambele abordări utilizează descrierile textuale ale articolelor, există o diferență în ceea ce privește definirea scorului de sentiment (înlocuirea ratingului numeric). Terzi et al. [22] calculează distanța dintre două cuvinte pe baza celei mai scurte distanțe dintre ele, în timp ce în abordarea propusă scorul de sentiment este obținut pe baza informațiilor derivate din Vader Lexicon [12].

Chiar dacă rezultatele cantitative din [22] sunt mai bune, abordarea prezentată este diferită din punct de vedere calitativ, utilizând o tehnică de filtrare colaborativă bazată pe lexicon. Tehnica propusă are valoare mai ales din punct de vedere semantic, luând în considerare polaritățile cuvintelor (pozitiv, negativ, neutru) în comparație cu [22], care se bazează pe setul de cuvinte comune. În general, această comparație evidențiază faptul că abordarea prezentată generează rezultate bune și confirmă încă o dată faptul că recenziile bazate pe text oferă într-adevăr informații valoroase pentru procesul de recomandare.

Rezultatele obținute în cadrul experimentelor numerice efectuate arată că abordarea prezentată poate fi utilizată cu succes pentru a rezolva sarcini de recomandare, pentru seturi de date care conțin recenzii ale utilizatorilor bazate pe text. Abordarea ar putea fi extinsă să în considerare și alte tipuri de elemente de recenzii în afară de cuvinte, cum ar fi subiectele recenziilor sau opiniile referitoare la aspecte.

5.3 Un Sistem de Recomandare Nesupervizat bazat pe Subiecte

Această secțiune prezintă un sistem de recomandare pentru articole New York Times (NYT RS), creat cu scopul de a propune cititorilor articole relevante pe anumite subiecte de interes. Articolele colectate de pe site-ul New York Times[1] sunt grupate pe baza subiectelor detectate manual cu ajutorul algoritmului de clustering K-Means. Clusterele rezultate sunt utilizate în procesul de recomandare pentru a sugera articole din același grup cu articolul deja citit. Această metodă îmbunătățește calitatea tehnicii clasice de recomandare k Nearest Neighbors (kNN).

Abordarea propusă are ca scop creșterea calității recomandărilor de articole din New York Times. Sistemul definește o listă de recomandări pentru un articol citit folosind o perspectivă nesupervizată bazată pe subiecte. Figura 5.3 prezintă arhitectura sistemului NYT RS construit. Există două componente principale ale sistemului: procesul de clustering și cel de recomandare. După faza de colectare a datelor, datele de intrare sunt trecute printr-o etapă de preprocesare. În continuare, este necesară o etapă de reprezentare a datelor, deoarece algoritmii de clusterizare pot trata doar date numerice. Algoritmul de clusterizare k-Means este utilizat pentru a obține grupuri de articole similare pe baza subiectelor. Indicele Silhouette [16] și indicele Dunn [10] sunt calculate pentru a evalua procesul de clusterizare. Cea de-a doua componentă principală a sistemului este algoritmul KNN bazat pe conținut [4] care determină cele mai similare k articole pentru un anumit articol citit de utilizator pe portalul New York Times, pe baza clusterelor bazate pe subiecte determinate anterior. În cele din urmă, calitatea recomandărilor generate este evaluată cu ajutorul măsurii acurateței, iar rezultatul este reprezentat de o listă de articole similare pe care un utilizator le poate citi pe un anumit subiect de interes.

Experimente Numerice

Mai multe configurații experimentale au fost concepute pentru a valida metodologia prezentată și pentru a determina calitatea sistemului de recomandare nesuperviyat, bazat pe subiecte.

În cadrul experimentelor efectuate au fost utilizate două seturi de date. Primul este format din 16.787 de articole de pe portalul New York Times. Acesta conține următoarele caracteristici:

- secțiunea este categoria articolului (de exemplu, politică, știință, jocuri etc.);
- material este tipul de articol (de exemplu, editorial, știri);
- titlu;
- abstract;
- data publicării;
- cuvinte-cheie;

Deoarece întregul text nu este public, se analizează doar rezumatul fiecărui articol. Sunt luate în considerare articolele între 1 ianuarie si 31 decembrie 2020. Pe baza unei analize (luând în considerare, de asemenea, câmpurile secțiune și cuvinte-cheie), au fost detectate cinci subiecte majore pentru datele

[1] https://www.nytimes.com

colectate: COVID-19, Donald Trump, protestele și mișcarea Black Lives Matter, Joe Biden și altele (precum incendiile de pădure, premiile Oscar etc.).

Al doilea set de date[2] este format din 10.732 de articole din New York Times de la sfârșitul anului 2017 și mijlocul anului 2018, cu următoarele caracteristici: autorul, titlul, conținutul, data publicării și adresa URL a articolului.

Procesul de recomandare se realizează direct pentru conținutul articolului. În plus, au fost detectate patru teme esențiale: mișcarea #metoo, Donald Trump, evenimente legate de Iran și diverse (de exemplu, schimbările climatice).

Procesul de clusterizare

Se aplică algoritmul k-Means bazat pe distanța euclidiană cu mai multe valori pentru parametrul k. Rezultatele experimentale arată că numărul optim de clustere corespunde numărului de subiecți identificați în cele două seturi de date. Prin urmare, pentru articolele New York Times din 2020, k este stabilit la 5, iar pentru articolele New York Times din 2017/2018, la 4. Rezultatele reflectă faptul că algoritmul k-Means se comportă foarte bine pe articolele New York Times pentru valorile k considerate.

Procesul de recomandare

Clusterele de articole sunt utilizate ca date de intrare în sistemul de recomandare bazat pe conținut KNN. Sunt definite mai multe scenarii experimentale pe baza măsurii de similaritate selectate (distanța cosinus, Jaccard sau euclidiană) și a diferitelor valori ale lui k (numărul de vecini).

Experimentele evidențiază următoarele aspecte:

- Cele mai bune valori ale acurateței au fost obținute folosind similaritatea Jaccard, deoarece aceasta este aplicată direct pe datele de intrare textuale, fără încorporări de cuvinte. Pe de altă parte, similaritatea Cosinus produce, de asemenea, rezultate bune care întăresc ideea că sistemul este performant.

- Cea mai relevantă valoare pentru k este cinci. Cu cât este mai mare numărul de vecini selectat k, cu atât calitatea recomandărilor este mai scăzută.

- Pentru setul de date cu articole NYT 2017/2018, sunt generate recomandări mai bune, deoarece este luat în considerare întregul articol, în comparație cu cel cu articole NYT 2020, în care sistemul primește ca intrare doar rezumatele.

Sistemul de recomandări pentru New York Times este un instrument care oferă utilizatorilor posibilitatea de a citi articole corelate cu interesele și așteptările lor. Partea originală este cea nesupervizată, care sporește calitatea recomandărilor. Sugerarea de articole înrudite cu cel care tocmai a fost citit, pe baza apartenenței lor la un cluster, se dovedește a fi o cale excelentă pentru explorare ulterioară. Acest aspect este susținut de rezultatele experimentale care arată valori ridicate pentru precizie. În plus, procesul de clustering în sine este evaluat pe baza indicilor Silhouette și Dunn.

Abordarea propusă reprezintă punctul de plecare al perspectivelor nesupervizate pentru sistemele de recomandare, dar sunt necesare investigații suplimentare. Planul implică utilizarea unor seturi

[2]https://www.kaggle.com/mathurinache/10700-articles-from-new-york-times

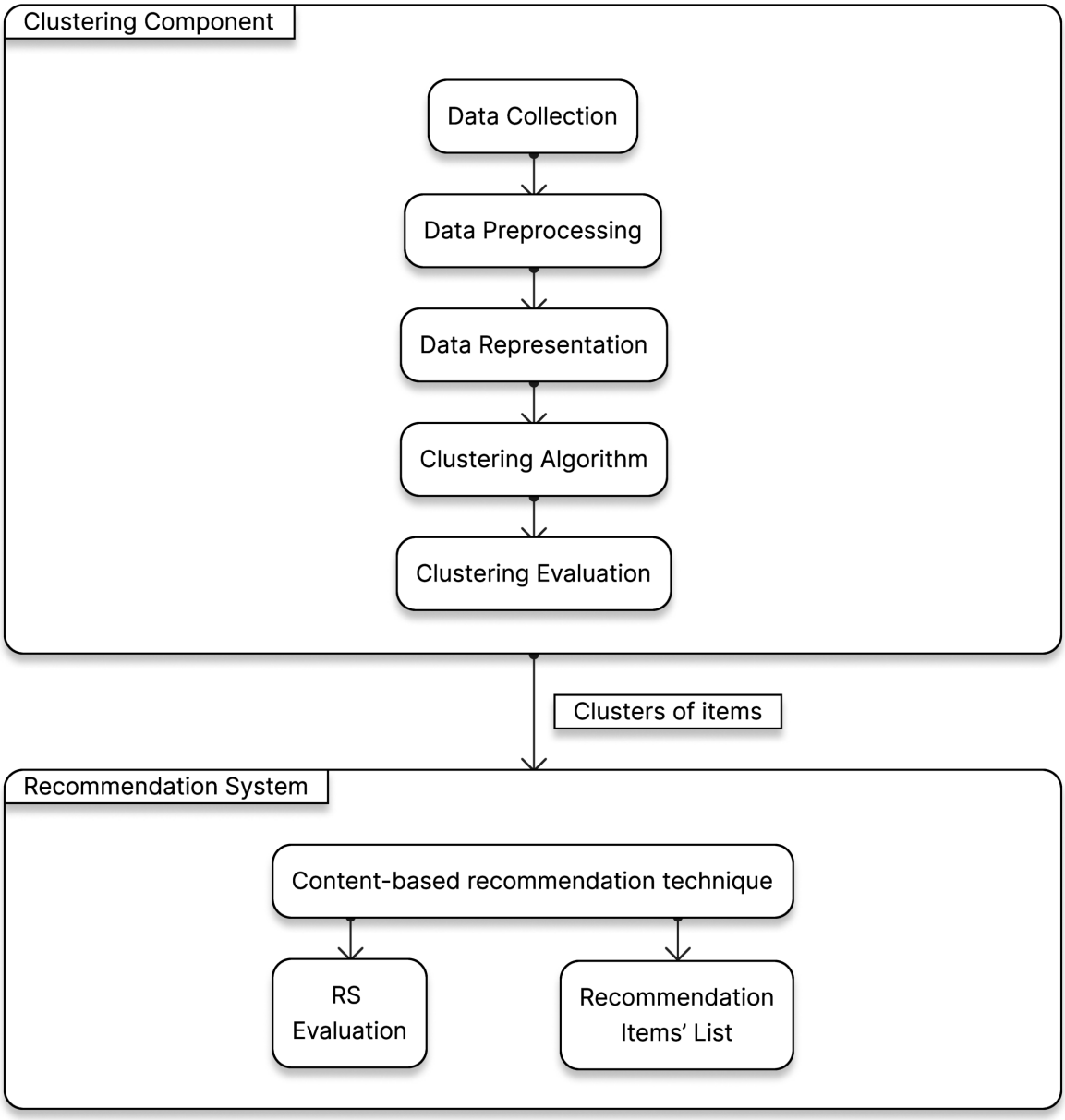

Figura 5.3: NYT Recommendation System Architecture.

de date multiple și mai extinse, deoarece abordarea prezentată poate fi aplicată oricărui set de date bazat pe text. În plus, ar trebui analizat în continuare impactul mai multor măsuri de similaritate asupra componentei de clusterizare și varietatea de încorporări de cuvinte care pot fi utilizate pentru a modela informațiile textuale.

Capitolul 6

Concluzii şi Extinderi

Acest capitol final trece în revistă concluziile generale ale tezei precum şi posibilităţile de extindere ulterioară.

Prima perspectivă [6], prezentată în capitolul 3, analizează cu atenție tehnica de filtrare colaborativă bazată pe memorie, cu un accent deosebit pe rolul critic al măsurilor de similaritate. Prin intermediul unor experimente extinse pe seturile de date MovieLens 1M și DataFiniti Hotel Reviews, studiul a dezvăluit care sunt măsurile de similaritate optime care trebuie utilizate în diferite contexte (de exemplu, luând în considerare dimensiunea și sparsitatea setului de date). Următoarele măsuri de similaritate sunt considerate în procesul de recomandare: Coeficientul de corelație Pearson, corelația Pearson constrânsă, similaritatea Cosinus, Cosinus ajustat, distanța euclidiană, corelația de rang Spearman, Jaccard și proximitate-impact-popularitate (PIP) [6]. PIP s-a dovedit a fi potrivită pentru CF bazată pe utilizator în seturi mari de date cu o sparsitate mai mică, în timp ce corelația de rang Spearman generează valori bune în scenariile CF bazate pe elemente. Similaritatea Jaccard a avut cele mai bune rezultate pentru seturi de date mai mici cu o sparsitate ridicată. În plus, alegerea unui număr mai mare de vecini (k) în algoritmul KNN sporește calitatea recomandărilor pentru seturi de date mari, în timp ce o valoare mai mică este preferabilă atunci când se lucrează cu un set de date redus.

În concluzie, capitolul 3 nu numai că explorează nuanțele măsurilor de similaritate, dar identifică și impactul acestora asupra acurateței recomandărilor, punând bazele pentru capitolele următoare.

Având în vedere concluziile prezentate de analiza comparativă a modului în care măsurile de similaritate influențează procesul de recomandare prezentate în capitolul 3 [6], Capitolul 4 introduce o nouă măsură de similaritate bazată pe sentiment, Atractivitate-Relevanţă-Popularity (ARP), cu scopul de a îmbunătăți performanța CF prin valorificarea recenziilor textuale, în locul celor numerice. Măsura ARP utilizează un lexicon de analiză a sentimentelor (Senti Word Net) pentru a extrage scorul de sentiment pentru o anumită recenzie bazată pe text. Astfel, evaluarea originală de la unu la cinci este înlocuită cu scorul de sentiment. Apoi, setul de date îmbogățit cu scorurile de sentiment este transmis algoritmului KNN și se generează o listă de n recomandări. Experimentele numerice sunt efectuate pe seturile de date Yelp Restaurants Review și Datafiniti Hotel Reviews, iar rezultatele, evaluate în termeni de MAE și RMSE, arată că ARP are performanțe mai bune decât majoritatea măsurilor clasice de similaritate, fiind potrivit pentru a fi utilizat pentru seturi de date care conțin doar recenzii bazate pe text. Acesta este un neajuns al măsurilor de similaritate tradiționale.

În plus, capitolul 4 propune un cadru de validare pentru măsura de similaritate ARP, bazat pe cinci

componente: utilitatea, expresivitatea, corelația cu alte măsuri, verificarea condițiilor unei metrici și robustețea la zgomot. Proiectarea cadrului de validare vizează revoluționarea procesului de dezvoltare de noi măsuri de similaritate prin evidențierea clară a valorii adăugate, comparativ cu modul în care sunt validate în prezent măsurile de similaritate în abordările din literatura de specialitate bazate exclusiv pe experimente numerice și pe metrici de evaluare precum acuratețea, precizia sau MAE.

Capitolul 5 este format din trei secțiuni (5.1, 5.2 și 5.3) care prezintă trei abordări individuale, având ca scop optimizarea diferitelor tehnici de recomandare.

În prima abordare [19], prezentată în secțiunea 5.1, tehnicile SA au fost utilizate împreună cu tehnica de filtrare colaborativă bazată pe utilizator (algoritmul KNN), prezentând îmbunătățiri remarcabile în ceea ce privește precizia și calitatea recomandărilor. Clasificatorul de sentimente primește ca intrare un set de date care conține atât recenzii numerice, cât și recenzii bazate pe text și produce ca ieșire un set de date etichetate pozitiv/negativ (fiecare evaluare numerică este îmbunătățită cu o etichetă de sentiment). Setul de date etichetate este transmis algoritmului KNN care generează o listă de recomandări. Pentru a evalua RS propus, au fost efectuate experimente numerice pe setul de date Yelp Restaurants' Reviews, iar rezultatele au fost comparate cu o abordare de bază (care nu ia în considerare etichetele de sentiment, ci doar evaluările numerice). Concluzia a fost că tehnicile de SA incluse în etapa de preprocesare a datelor din cadrul procesului de recomandare sporesc performanța algoritmului KNN și calitatea sugestiilor în ceea ce privește precizia, recall, f-score și MAE.

Abordarea [7], descrisă în secțiunea 5.2, introduce o tehnică de filtrare colaborativă KNN bazată pe lexicon. Folosindu-se de lexiconul Vader pentru determinarea ratingurilor de sentiment, abordarea a demonstrat succesul în sarcinile de recomandare cu seturi de date care conțin recenzii ale utilizatorilor bazate pe text. Această secțiune nu numai că a evidențiat succesul abordării bazate pe lexicon în optimizarea algoritmului KNN, dar a sugerat, de asemenea, lucrări viitoare care implică luarea în considerare a diferitelor elemente de recenzie în afară de cuvinte.

Sistemul de recomandare al New York Times prezentat în secțiunea 5.3, aduce o particularitate nesupervizată. Algoritmul K-Means definește clusterele în funcție de cele mai frecvente subiecte din articolele New York Times. Clusterele de articole rezultate sunt utilizate ca intrare în algoritmul de filtrare colaborativă KNN, iar recomandările sunt generate din clusterul din care face parte articolul citit în acel moment. Evaluarea procesului de clusterizare, luând în considerare indicii Silhouette și Dunn, a validat și mai mult abordarea propusă. Lucrările viitoare includ explorarea diferitelor măsuri de similaritate și încorporări de cuvinte, analizarea impactului acestora asupra componentelor de clusterizare și extinderea abordării propuse prin folosirea unor seturi de date multiple.

În mod colectiv, această teză a aprofundat diverse fațete ale sistemelor de recomandare, de la complexitatea măsurilor de similaritate la analiza sentimentelor, abordări bazate pe lexicon și sisteme de recomandare nesupervizate bazate pe subiecte. Constatările și contribuțiile stabilesc o bază solidă pentru avansarea cercetării în domeniul sistemelor de recomandare. În concluzie, teza nu numai că a contribuit cu informații valoroase la domeniul sistemelor de recomandare, dar a și pregătit terenul pentru explorări și perfecționări viitoare.

Bibliografie

[1] Ajay Agarwal and Minakshi Chauhan. Similarity measures used in recommender systems: a study. *International Journal of Engineering Technology Science and Research IJETSR, ISSN*, pages 2394–3386, 2017.

[2] Hyung Jun Ahn. A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem. *Information Sciences*, 178(1):37–51, 2008.

[3] Jesus Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutierrez. Recommender systems survey. *Knowledge-based systems*, 46:109–132, 2013.

[4] Bei-Bei Cui. Design and implementation of movie recommendation system based on knn collaborative filtering algorithm. In *ITM web of conferences*, volume 12, page 04008. EDP Sciences, 2017.

[5] Rafael M D'Addio and Marcelo G Manzato. A sentiment-based item description approach for knn collaborative filtering. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, pages 1060–1065, 2015.

[6] Mara Deac-Petruşel. A comparative analysis of similarity measures in memory-based collaborative filtering. In *Artificial Intelligence and Soft Computing: 19th International Conference, ICAISC 2020, Zakopane, Poland, October 12-14, 2020, Proceedings, Part II 19*, pages 140–151. Springer, 2020.

[7] Mara Deac-Petruşel. A lexicon-based collaborative filtering approach for recommendation systems. In *International Conference on Agents and Artificial Intelligence (ICAART)*, pages 203–210, 2022.

[8] Mara Deac-Petruşel and Sergiu Limboi. A sentiment-based similarity model for recommendation systems. In *2020 22nd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, pages 224–230. IEEE, 2020.

[9] Ruihai Dong, Michael P O'Mahony, Markus Schaal, Kevin McCarthy, and Barry Smyth. Sentimental product recommendation. In *Proceedings of the 7th ACM Conference on Recommender Systems*, pages 411–414, 2013.

[10] Tanvi Gupta and Supriya P Panda. Clustering validation of clara and k-means using silhouette & dunn measures on iris dataset. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pages 10–13. IEEE, 2019.

[11] Davide Feltoni Gurini, Fabio Gasparetti, Alessandro Micarelli, and Giuseppe Sansonetti. A sentiment-based approach to twitter user recommendation. *RSWeb@ RecSys*, 1066, 2013.

[12] Clayton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, 2014.

[13] C. Jiang, L. Xia, and S. Li. A sentiment-based similarity method for cold-start recommendations. *Knowledge-Based Systems*, 189:105116, 2020.

[14] Q. Li, M. Zhang, and L. Li. A collaborative filtering recommendation algorithm based on sentiment similarity. *Mathematical Problems in Engineering*, 2019:1–9, 2019.

[15] Sergiu Limboi and Mara Deac-Petruşel. A validation framework for arp similarity measure. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1266–1271, 2021.

[16] A Rasid Mamat, F Susilawati Mohamed, M Afendee Mohamed, N Mohd Rawi, and M Isa Awang. Silhouette index for determining optimal k-means clustering on images in different color models. *Int. J. Eng. Technol*, 7(2):105–109, 2018.

[17] NA Osman, Shahrul Azman Mohd Noah, and M Darwich. Contextual sentiment based recommender system to provide recommendation in the electronic products domain. *International Journal of Machine Learning and Computing*, 9(4):425–431, 2019.

[18] Mara Petruşel. An unsupervised topic-driven new york times recommendation system. In *2022 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, pages 1–6. IEEE, 2022.

[19] Mara Petruşel and Sergiu-George Limboi. A restaurants recommendation system: Improving rating predictions using sentiment analysis. In *2019 21st International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, pages 190–197. IEEE, 2019.

[20] Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor. *Introduction to Recommender Systems Handbook*, pages 1–35. Springer US, Boston, MA, 2011.

[21] Mr Sridhar Dilip Sondur, Mr Amit P Chigadani, and Shantharam Nayak. Similarity measures for recommender systems: a comparative study. *Journal for Research*, 2(3), 2016.

[22] Maria Terzi, Matthew Rowe, Maria-Angela Ferrario, and Jon Whittle. Text-based user-knn: Measuring user similarity based on text reviews. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 195–206. Springer, 2014.