(19) OFICIUL DE STAT
PENTRU INVENŢII ŞI MĂRCI
Bucureşti

ROMÂNIA

(11) **RO 134064 A2**
(51) **Int.Cl.**
G06F 16/30 (2019.01)

(12) # CERERE DE BREVET DE INVENŢIE

(21) Nr. cerere: **a 2018 00606**

(22) Data de depozit: **24/08/2018**

(41) Data publicării cererii:
30/04/2020 BOPI nr. **4/2020**

(71) Solicitant:
• **RESEARCH TECHNOLOGY S.R.L.**,
ŞOS.VIRTUŢII, NR.19D, ET.6, SECTOR 6, BUCUREŞTI, B, RO

(72) Inventatori:
• **RUŞEŢI STEFAN**, *STR.DOAMNA GHICA, NR.32B, BL.T3, AP.608, SECTOR 2, BUCUREŞTI, RO;*
• **DASCĂLU MIHAI**, *BD.ION MIHALACHE, NR.126, BL.1, SC.A, ET.5, AP.37, SECTOR 1, BUCUREŞTI, B, RO;*
• **PARASCHIV IONUŢ**, *STR.MAŞINA DE PÂINE, NR.20, BL.OD37, SC.3, ET.3, AP.103, SECTOR 2, BUCUREŞTI, B, RO;*
• **BĂNICĂ KARL COSMIN**, *STR.BLÂNDEŞTI, NR.24C, SECTOR 4, BUCUREŞTI, B, RO;*
• **MIHAI CORINA DANIELA**, *DRUMUL SĂRII, NR.18, BL.A41, SC.A, ET.4, AP.15, SECTOR 6, BUCUREŞTI, B, RO*

Această publicaţie include şi modificările descrierii, revendicărilor şi desenelor, depuse conform art. 35, alin. (20), din HG nr. 547/2008.

(54) # METODĂ ŞI PROCES DE RECOMANDARE DE ARTICOLE ŞI AUTORI RELEVANŢI UTILIZÂND UN SPAŢIU VECTORIAL DE REPREZENTARE UNIFICATĂ

(57) **Rezumat:**

Invenţia se referă la o metodă de recomandare de articole şi autori relevanţi, utilizând un spaţiu vectorial de reprezentare unificată, rezultat din antrenarea unei reţele (**1**) neurale pe un corpus de articole. În vederea calculului reprezentării unificate în spaţiul vectorial, reţeaua (**1**) foloseşte descrierea articolului, autorii acestuia, cât şi articolele citate, reprezentarea fiind folosită atât pentru ordonarea după importanţă a autorilor şi articolelor dintr-un corpus (**2**), cât şi pentru recomandarea (**3**) de articole similare relevante. Pentru calculul importanţei unui articol dintr-un domeniu sunt folosite metrici de centralitate în spaţiul nou generat, iar în cazul documentelor noi este folosită reţeaua (**1**) neurală deja antrenată pe un corpus dat, în vederea generării unei reprezentări vectoriale în acelaşi spaţiu unificat în care se calculează similaritatea cu celelalte documente din corpus, şi se sortează descrescător după proximitate.

Revendicări iniţiale: 4
Revendicări amendate: 2
Figuri: 1

RO 134064 A2

Descriere tehnică

Invenția se referă la o metodă de învățare a unor reprezentări vectoriale unificate a articolelor și a autorilor. Acest model este inspirat din alte modele de reprezentare a cuvintelor, care, pentru calculul acesteia, învață să prezică cuvintele care apar în contexte similare cu un termen dat.

Brevetul US 4839853A descrie un model de reprezentare a unui document într-un spațiu vectorial latent pornind de la o matrice termen-document de apariții ale cuvintelor în documente. Similaritatea cosinus este aplicată între vectorii documentelor în vederea calculului distanței semantice dintre acestea. Dezavantajul metodei constă în faptul că aceasta ține cont doar de conținutul documentelor, și nu de alte relații alternative care pot fi ușor identificate între documente (spre exemplu, autori comuni care induc interese similare, sau referințe bibliografice identice care denotă surse de interes comun).

Brevetul US 6523026B1 prezintă un model computațional ce constă în atașarea unor vectori multidimensionali documentelor text cu scopul de a le grupa în categorii. Reprezentările abstracte ale termenilor sunt puncte într-un spațiu vectorial, încapsulând tiparele apariției termenilor din domeniul sursă. Dezavantajul principal al acestei soluții este faptul că recomandarea nu se realizează într-un mod suficient de granular, utilizatorul primind doar un grup de documente similare.

Brevetul US 6922699B2 este similar cu cel anterior, însă, în plus, folosește în cadrul vectorilor multidimensionali o componentă ce ține cont de comportamentul utilizatorilor, prin agregarea istoricului de navigare al acestora. Modelul calculează distanțe semantice între documente cu scopul de a le asocia categorii, folosite ulterior pentru recomandare. Dezavantajul acestei metode este faptul că

nu integrează metrici de similaritate semantică pentru calcularea distanțelor dintre documente din prisma conținutului acestora.

Brevetul CN 103617157A prezintă un model de aproximare a distanțelor semantice dintre documente folosind cuvintele cheie din cadrul acestora. Distanțele se calculează individual și se adună pentru generarea unui scor global pentru aproximarea distanței dintre două documente. Dezavantajul principal al metodei este faptul că nu ține cont de relațiile alternative menționate anterior pentru exprimarea similarității dintre documente.

Avantajele reprezentării unificate conform invenției constau în:

- Reprezentarea unui articol ține cont de descrierea textuală aferentă, autorii acestuia și articolele citate / referite;
- Metoda permite calculul similarității între autori, nu doar între documente;
- Două articole pot fi mai apropiate în spațiul de reprezentare unificată inclusiv în condițiile în care acestea au autori sau articole citate similare, nu neapărat aceleași;
- Viteză în obținerea reprezentării unificate prin folosirea unei arhitecturii optimizate de rețele neurale versus descompunerile clasice bazate pe valori singulare (SVD - Singular Value Decomposition).

Modelul propus, prezentat în **Error! Reference source not found.**, folosește textul din descrierea articolului, autorii și articolele referite pentru a prezice articolele similare cu acesta. Descrierea articolului (rezumat, fragment relevant sau textul integral aferent) este împărțită în cuvinte, reprezentate ulterior într-un spațiu semantic pre-antrenat (**1.1**), de dimensiune d_c. Autorii articolului, precum și articolele citate sunt reprezentați prin vectori (de dimensiuni d_{aut} și d_{art}) antrenați pe un corpus de articole dintr-un domeniu, inițializați în mod aleatoriu (**1.2** și **1.3**).

În urma acestor preprocesări, cele trei intrări ale rețelei sunt reprezentate prin trei matrice de dimensiuni variabile , care depind de numărul de cuvinte din descriere

(n_c), numărul de autori al articolului (n_{aut}), respectiv numărul de articole citate (n_{art}). Aceste reprezentări diferite sunt aduse într-o dimensiune fixă (spre exemplu, prin aplicarea mediei aritmetice sau ponderare printr-un model bazat pe atenție) (**1.4**, **1.5** și **1.6**) și apoi concatenate, obținând un vector de dimensiune cumulată $d_c + d_{aut} + d_{art}$ (**1.7**). O transformare vectorială (**1.8**) este aplicată asupra acestui vector pentru a-l transpune în spațiul vectorial al articolelor, unde va fi comparat cu toate articolele din corpus (**1.9**).

Pentru calculul similarității dintre articole se folosește produsul scalar dintre vectorul construit de rețea și vectorii fiecărui articol din corpus, aceiași vectori utilizați totodată ca intrare în rețea. În faza de antrenare, modelul învață să detecteze articolele similare, pentru fiecare articol din corpus. Un articol este considerat similar dacă respectă una dintre următoarele condiții:

- Similaritatea semantică dintre descrieri, calculată conform unui model semantic, depășește un prag prestabilit;
- Cele două articole au cel puțin un autor în comun;
- Cele două articole sunt împreună citate de un alt articol;
- Cele două articole citează un același alt articol.

În urma etapei de antrenare, rețeaua învață reprezentări vectoriale unificate pentru fiecare articol și autor din corpus, precum și parametrii necesari pentru transformarea vectorială din ultimul strat al modelului. Aceste reprezentări și parametri sunt folosiți ulterior de cele două module propuse, *Modulul de ordonare a articolelor după importanță* (**2**) și *Modulul de recomandare de articole relevante* (**3**).

Modulul de ordonare a articolelor după importanță (**2**) folosește vectorii calculați prin metoda descrisă anterior pentru a construi un graf bi-modal neorientat conținând 2 tipuri de noduri (articole și autori) pe care se pot calcula diverse metrici de centralitate. Graful corespunzător colecției de articole și autori aferenți utilizați la antrenarea modelului are la bază muchii ce reflectă o similaritate

cosinus în spațiul de reprezentare unificată peste un prag minim prestabilit în intervalul [0; 1]. Diverse metrici de centralitate specifice teoriei grafurilor (spre exemplu, gradul unui nod reflectat în suma ponderilor muchiilor, centralitatea ca inversul distanțelor minime la toate nodurile din graf) sunt aplicate ulterior construirii acestui graf, în vederea identificării importanței unui nod în cadrul colecției de documente.

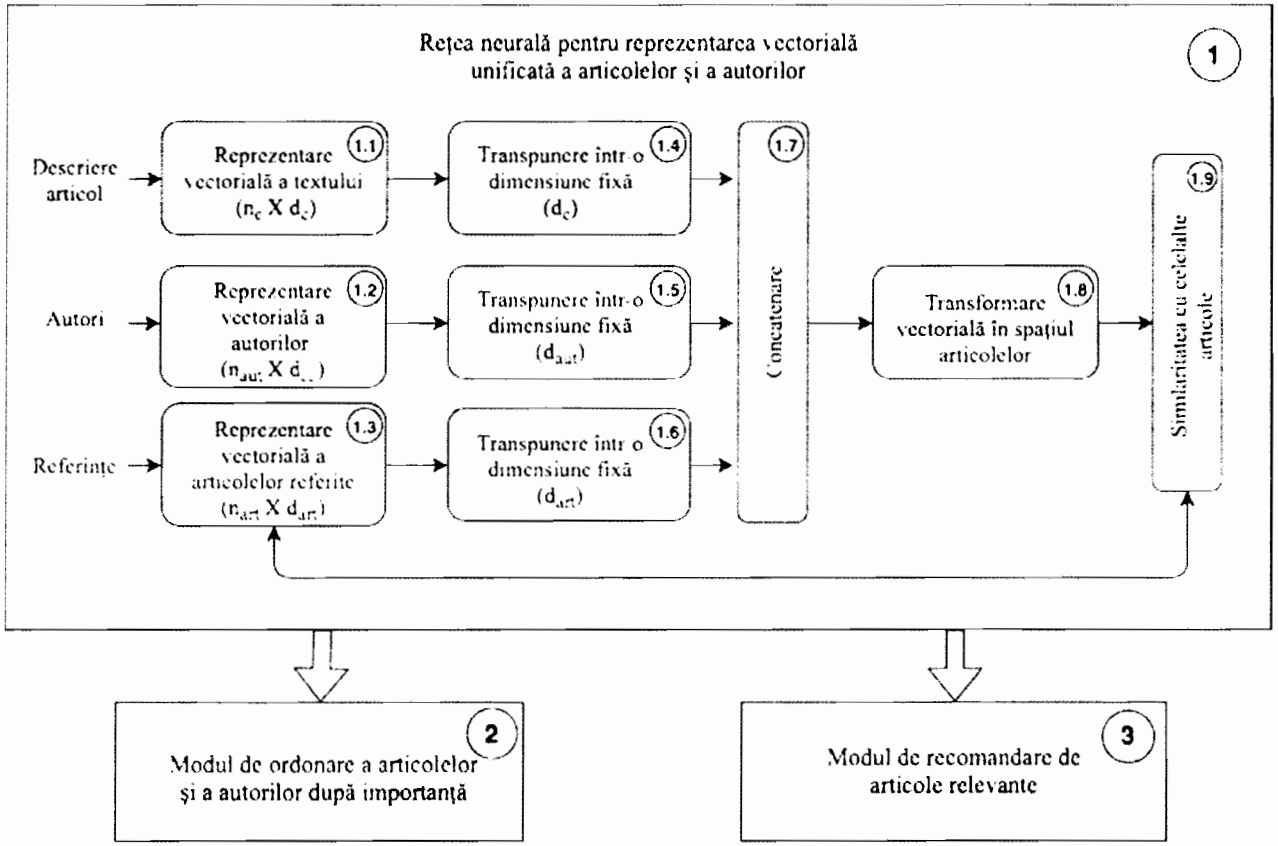
Modulul de recomandare de articole relevante **(3)** este utilizat pentru reprezentarea unor articole noi, neutilizate în faza de antrenare. Aceste articole sunt trecute prin aceeași rețea neurală antrenată anterior, folosind descrierea și reprezentarea cuvintelor conținute, precum și autorii și articolele referite care existau în corpusul inițial utilizat la antrenare și pentru care există vectori calculați. Rețeaua calculează un vector pentru articolul nou, care poate fi folosit pentru a găsi articole similare utilizând proximitatea sau similaritatea cosinus din spațiul vectorial.

Revendicări

La invenția „Metodă de recomandare de articole și autori relevanți utilizând un spațiu vectorial de reprezentare unificat" revendicăm:

1. Metoda de calcul a reprezentării unificate a articolelor și autorilor dintr-un domeniu într-un spațiu vectorial, în vederea facilitării regăsirii de resurse relevante.
2. Arhitectura specifică de rețele neurale (**1**) care optimizează calculul similarității dintre articole având la bază 3 tipuri de legături între documente: co-autori (autori comuni), co-citări (articolele referite identice), și similaritatea semantică între textele aferente (abstracte, fragmente de text sau documentul integral).
3. Metoda de ordonare a articolelor și a autorilor (**2**) folosind metrici de centralitate în spațiul vectorial de reprezentare unificată.
4. Metoda de recomandare de articole relevante (**3**) pornind de la un nou document folosind metoda de calcul a reprezentării acestuia în spațiul vectorial unificat antrenat pentru un corpus dat, precum și funcția de similaritate aplicată în spațiul de reprezentare unificată aferent.

Figura 1 - Schema bloc a metodei de recomandare de articole şi autori relevanţi utilizând un spaţiu vectorial de reprezentare unificată



n_c număr cuvinte;

n_{aut} număr autori document;

n_{art} număr articole referite

d_c dimensiunea vectorului de reprezentare aferent unui cuvânt;

d_{aut} dimensiunea vectorului de reprezentare al unui autor;

d_{art} dimensiunea vectorului de reprezentare unificată aferent unui articol;

a 2018 00606 24/08/2018

45

Descriere tehnică

Invenția se referă la o metodă de învățare a unor reprezentări vectoriale unificate a articolelor și a autorilor. Acest model este inspirat din alte modele de reprezentare a cuvintelor, care, pentru calculul acesteia, învață să prezică cuvintele care apar în contexte similare cu un termen dat.

Brevetul US 4839853A descrie un model de reprezentare a unui document într-un spațiu vectorial latent pornind de la o matrice termen-document de apariții ale cuvintelor în documente. Similaritatea cosinus este aplicată între vectorii documentelor în vederea calculului distanței semantice dintre acestea. Dezavantajul metodei constă în faptul că aceasta ține cont doar de conținutul documentelor, și nu de alte relații alternative care pot fi ușor identificate între documente (spre exemplu, autori comuni care induc interese similare, sau referințe bibliografice identice care denotă surse de interes comun).

Brevetul US 6523026B1 prezintă un model computațional ce constă în atașarea unor vectori multidimensionali documentelor text cu scopul de a le grupa în categorii. Reprezentările abstracte ale termenilor sunt puncte într-un spațiu vectorial, încapsulând tiparele apariției termenilor din domeniul sursă. Dezavantajul principal al acestei soluții este faptul că recomandarea nu se realizează într-un mod suficient de granular, utilizatorul primind doar un grup de documente similare.

Brevetul US 6922699B2 este similar cu cel anterior, însă, în plus, folosește în cadrul vectorilor multidimensionali o componentă ce ține cont de comportamentul utilizatorilor, prin agregarea istoricului de navigare al acestora. Modelul calculează distanțe semantice între documente cu scopul de a le asocia categorii, folosite ulterior pentru recomandare. Dezavantajul acestei metode este faptul că nu integrează metrici de similaritate semantică pentru calcularea distanțelor dintre documente din prisma conținutului acestora.

Brevetul CN 103617157A prezintă un model de aproximare a distanțelor semantice dintre documente folosind cuvintele cheie din cadrul acestora. Distanțele se

calculează individual și se adună pentru generarea unui scor global pentru aproximarea distanței dintre două documente. Dezavantajul principal al metodei este faptul că nu ține cont de relațiile alternative menționate anterior pentru exprimarea similarității dintre documente.

Avantajele reprezentării unificate conform invenției constau în:

- Reprezentarea unui articol ține cont de descrierea textuală aferentă, autorii acestuia și articolele citate / referite;
- Metoda permite calculul similarității între autori, nu doar între documente;
- Două articole pot fi mai apropiate în spațiul de reprezentare unificată inclusiv în condițiile în care acestea au autori sau articole citate similare, nu neapărat aceleași;
- Viteză în obținerea reprezentării unificate prin folosirea unei arhitecturii optimizate de rețele neurale versus descompunerile clasice bazate pe valori singulare (SVD - Singular Value Decomposition).

Modelul propus, prezentat în Figura 1, folosește textul din descrierea articolului, autorii și articolele referite pentru a prezice articolele similare cu acesta. Descrierea articolului (rezumat, fragment relevant sau textul integral aferent) este împărțită în cuvinte, reprezentate ulterior într-un spațiu semantic pre-antrenat (**1.1**), de dimensiune d_c. Autorii articolului, precum și articolele citate sunt reprezentați prin vectori (de dimensiuni d_{aut} și d_{art}) antrenați pe un corpus de articole dintr-un domeniu, inițializați în mod aleatoriu (**1.2** și **1.3**).

În urma acestor preprocesări, cele trei intrări ale rețelei sunt reprezentate prin trei matrice de dimensiuni variabile, care depind de numărul de cuvinte din descriere (n_c), numărul de autori al articolului (n_{aut}), respectiv numărul de articole citate (n_{art}). Aceste reprezentări diferite sunt aduse într-o dimensiune fixă (spre exemplu, prin aplicarea mediei aritmetice sau ponderare printr-un model bazat pe atenție) (**1.4**, **1.5** și **1.6**) și apoi concatenate, obținând un vector de dimensiune cumulată $d_c + d_{aut} + d_{art}$ (**1.7**). O transformare vectorială (**1.8**) este aplicată asupra acestui vector pentru a-l transpune în spațiul vectorial al articolelor, unde va fi comparat cu toate articolele din corpus (**1.9**).

Pentru calculul similarității dintre articole se folosește produsul scalar dintre vectorul construit de rețea și vectorii fiecărui articol din corpus, aceiași vectori utilizați totodată ca intrare în rețea. În faza de antrenare, modelul învață să detecteze articolele similare, pentru fiecare articol din corpus. Un articol este considerat similar dacă respectă una dintre următoarele condiții:

- Similaritatea semantică dintre descrieri, calculată conform unui model semantic, depășește un prag prestabilit;
- Cele două articole au cel puțin un autor în comun;
- Cele două articole sunt împreună citate de un alt articol;
- Cele două articole citează un același alt articol.

În urma etapei de antrenare, rețeaua învață reprezentări vectoriale unificate pentru fiecare articol și autor din corpus, precum și parametrii necesari pentru transformarea vectorială din ultimul strat al modelului. Aceste reprezentări și parametri sunt folosiți ulterior de cele două module propuse, *Modulul de ordonare a articolelor după importanță* (**2**) și *Modulul de recomandare de articole relevante* (**3**).

Modulul de ordonare a articolelor după importanță (**2**) folosește vectorii calculați prin metoda descrisă anterior pentru a construi un graf bi-modal neorientat conținând 2 tipuri de noduri (articole și autori) pe care se pot calcula diverse metrici de centralitate. Graful corespunzător colecției de articole și autori aferenți utilizați la antrenarea modelului are la bază muchii ce reflectă o similaritate cosinus în spațiul de reprezentare unificată peste un prag minim prestabilit în intervalul [0; 1]. Diverse metrici de centralitate specifice teoriei grafurilor (spre exemplu, gradul unui nod reflectat în suma ponderilor muchiilor, centralitatea ca inversul distanțelor minime la toate nodurile din graf) sunt aplicate ulterior construirii acestui graf, în vederea identificării importanței unui nod în cadrul colecției de documente.

Modulul de recomandare de articole relevante (**3**) este utilizat pentru reprezentarea unor articole noi, neutilizate în faza de antrenare. Aceste articole sunt trecute prin aceeași rețea neurală antrenată anterior, folosind descrierea și reprezentarea cuvintelor conținute, precum și autorii și articolele referite care existau în corpusul inițial utilizat la antrenare și pentru care există vectori calculați. Rețeaua calculează

un vector pentru articolul nou, care poate fi folosit pentru a găsi articole similare utilizând proximitatea sau similaritatea cosinus din spațiul vectorial.

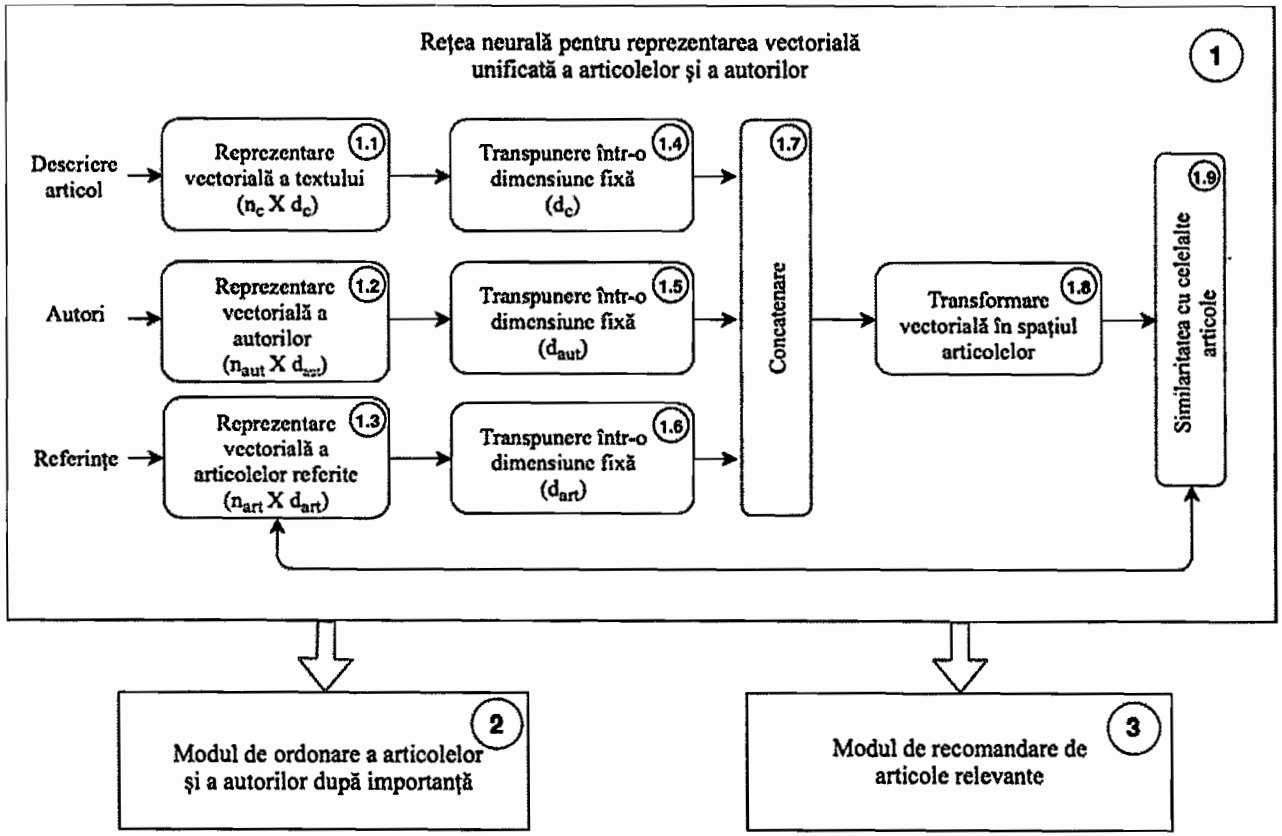
Revendicări

1. Metodă de recomandare de articole și autori relevanți utilizând un spațiu vectorial de reprezentare unificat, cuprinzând o fază de antrenare și o fază de test, **caracterizată prin aceea că**, în faza de *antrenare*, se aplică un algoritm bazat pe rețele neurale pentru generarea unei reprezentări unificate într-un spațiu vectorial al articolelor și autorilor dintr-un domeniu, algoritm care cuprinde o primă etapă în care se inițializează trei intrări ale rețelei, și anume matricea cu descrierea articolului (rezumat, fragment relevant sau textul integral aferent) care conține fiecare cuvânt și reprezentarea vectorială a acestuia într-un spațiu semantic pre-antrenat, matricea cu autorii ai căror vectori sunt inițializați în mod aleatoriu și matricea cu articolele referite ai căror vectori sunt inițializați în mod aleatoriu; într-o a două etapă, se transpun reprezentările diferite ale matricelor anterioare într-o formă fixă; într-o a treia etapă se concatenează reprezentările fixe ale textului, autorilor și articolelor referite; într-o a patra etapă, se aplică o transformare vectorială a reprezentării concatenate în spațiul articolelor; într-o a cincea etapă se compară articolul curent cu toate celelalte articole din corpus, ale căror reprezentări sunt folosite la intrare, prin utilizarea unui produs scalar care determină similaritatea dintre două articole; într-o a șasea etapă, se propagă eroarea în rețea, eroare care reflectă 4 condiții: similaritate semantică dintre descrieri, existența unui autor în comun, citarea celor două articole într-un alt articol, sau citarea unui același articol; într-o a șaptea etapă, se ordonează articolele și autorii, după finalizarea antrenării și convergența rețelei, folosind metrici de centralitate în spațiul vectorial de reprezentare unificată și se trece la faza de *test* în care, într-o primă etapă, se calculează reprezentarea unui nou document în spațiul vectorial unificat pentru un corpus dat folosind rețeaua neurală antrenată în prealabil și în a doua etapă se identifică cele mai apropiate articole și autori din spațiul vectorial existent folosind valori în

ordine descrescătoare a similarității aplicate în spațiul de reprezentare unificat aferent.

2. Metodă conform revendicării 1, caracterizată prin aceea că rețeaua neurală are o arhitectură recurentă.

Figura 1 - Schema bloc a metodei de recomandare de articole și autori relevanți utilizând un spațiu vectorial de reprezentare unificată



n_c număr cuvinte; d_c dimensiunea vectorului de reprezentare aferent unui cuvânt;

n_{aut} număr autori document; d_{aut} dimensiunea vectorului de reprezentare al unui autor;

n_{art} număr articole referite d_{art} dimensiunea vectorului de reprezentare unificată aferent unui articol;