IDENTIFICAREA ORIGINALITĂȚII UNEI LUCRĂRI DE SPECIALITATE FOLOSIND ANALIZA INTRINSECĂ A PLAGIATULUI

Mădălina Zurini

madalina.zurini@csie.ase.ro

Academia de Studii Economice

Rezumat: În cadrul articolului se prezintă conceptul de proprietate intelectuală în contextul publicării articolelor de specialitate. Nivelul de originalitate derivă din analiza dreptului de proprietate, fiind o componentă definitorie şi prezentată în antiteză cu conceptul de plagiat. Făcând o trecere la tipurile de analiză a plagiatului, analiza intrinsecă este exemplificată prin propunerea unei metrici de evaluare a stilului de scriere al unui autor din punct de vedere al bogăției vocabularului folosit în concordanță cu nivelul semantic abordat. Metrica propusă este testată folosind o bază de test formată din 17 lucrări de specialitate realizate de un autor pe parcursul a 13 ani. Interpretarea rezultatelor evidențiază avantajele aduse de completarea analizei cu evaluarea nivelului semantic.

Cuvinte cheie: stilometrie, plagiat, metrici, originalitate.

Abstract: Within the paper, the concept of intellectual property is presented in the context of publishing scientific papers. The level of originality derives from the analyses of intellectual property, a defining component presented in antithesis with the concept of plagiarism. In a further analyses, the main methods of plagiarism identification are presented, concentrating on the intrinsic plagiarism using a proposed metric for evaluating the writing style of an author in accordance to the semantic approach. The proposed metric is tested using a dataset formed out of 17 research articles conducted by an author over 13 years. The interpretation of the results highlights the advantages brought by adding a semantic layer evaluation to the current analysis.

Keywords: stylometry, plagiarism, metrics, originality.

1. Nivelul de originalitate şi dreptul de proprietate intelectuală

Cercetările privind dreptul de proprietate intelectuală se confruntă cu stabilirea nivelului de originalitate, în antiteză cu acțiunea de plagiat care este definită ca fiind însuşirea integrală sau parțială a ideilor, expresiilor, metodelor sau procedurilor altor autori şi prezentarea lor drept creație personală. În legile anglo-americane, argumentele de ordin economic şi cele care țin de politicile publice prevalează în elaborarea şi dezvoltarea legilor dreptului de proprietate intelectuală, în timp ce în spațiul european argumentele de ordin moral şi cele de drept civil fundamentează elaborarea aceloraşi legi.

Cadrul legislativ prezent nu rezolvă identificarea plagiatului şi a nivelului de originalitate al lucrărilor ştiințifice. Prezentul proiect îşi propune aplicarea legislativă a drepturilor de proprietate intelectuală în contextul publicării lucrărilor de specialitate.

În practică, există diferite tipuri de plagiat, cele mai des întâlnite fiind: plagiatul copy-paste, parafrazarea, plagiatul prin traducere din alte limbi, plagiatul artistic, al ideilor, al codului sursă şi nefolosirea corespunzătoare a citărilor. În [1], se prezintă faptul că plagiatul prin parafrazare este analizat, realizând o clasificare a tipurilor întâlnite, precum şi testarea produselor software de detectare a plagiatului la nivelul procentului de corectitudine a identificării parafrazării în cadrul unui document text.

Creativitatea, văzută ca o formă de originalitate, reprezintă capacitatea de a aduce ceva nou, original şi adecvat realității, definitorii pentru creativitate fiind noutatea şi originalitatea. Astfel, pentru a putea analiza nivelul de originalitate al unei lucrări ştiințifice, este nevoie de a crea o antiteză între această componentă de creativitate şi componenta de plagiat.

Pornind de la obiectele cu care se operează în cadrul prezentei cercetări, lucrări de specialitate realizate de autori străini şi români, se defineşte componenta de frază semantică ca acea componentă compactă din cadrul unei lucrări, formată din una sau mai multe fraze alăturate, care este semnificativ diferită de frazele semantice imediat anterioară şi ulterioară. A spune că o lucrare este originală este echivalent cu rezultatul evaluării acelei lucrări din punct de vedere al plagiatului.

Se defineşte metrica *IEO*, Indicator de Evaluare a Originalității, ca fiind raportul dintre numărul de fraze semantice originale din cadrul unei lucrări analizate raportat la numărul total de fraze semantice identificate în lucrare.

$$IEO = \frac{nr_{fraze\ originale}}{nfs}$$

unde:

- $nr_{fraze\ originale}$ reprezintă numărul total de fraze semantice originale din cadrul unei lucrări şi este egal cu

  $$Card(\{fs_i \forall i = \overline{1,nfs}, fs_i \neq fs_{i-1}, fs_i \neq fs_{i+1}, fp(fs_i) < \varepsilon\})$$

- *nfs* reprezintă numărul total de fraze semantice identificate şi este egal cu $Card(\{fs_i \forall i = \overline{1,nfs}, fs_i \neq fs_{i-1}, fs_i \neq fs_{i+1}\})$;

- *fp(x)* reprezintă funcția de evaluare a gradului de plagiat identificat în cadrul unei fraze semantice, având drept codomeniu intervalul [0,1];

- ε reprezintă pragul procentual maximal admis din punct de vedere al rezultatului funcției de evaluare a plagiatului unei fraze semantice, $\varepsilon \in [0,1]$.

Prezenta metrică este propusă în strânsă legătură cu definiția dată de Osiceanu în lucrarea sa privind originalitatea, [11]. Legea dreptului de autor subliniază că „originalitatea" semnifică în mod fundamental faptul că o lucrare provine din inspirația autorului şi că nu a fost copiată din altă sursă. Prin urmare, „original" este folosit în sensul de originar, în scopul de a identifica sursa în care o lucrare îşi are originea.

Cu cât o lucrare conține mai puține fraze care se confundă cu cercetările anterioare, cu atât acea lucrare va avea un grad de originalitate mai ridicat. Lucrarea de față foloseşte conceptul de plagiat nu doar la nivelul restrâns şi foarte cunoscut al acestuia, în sensul de copiat fără a referi corect, moral şi legal sursa textului, ci şi într-un sens al ideii, al direcției de cercetare care poate influența cercetarea unui autor ținând cont şi de cercetările anterioare similare ale altor autori. O lucrare este originală atunci când tratează un concept, domeniu, situație nouă sau existentă într-o manieră unică față de celelalte cercetări anterioare.

Prezentele abordări ale identificării plagiatului cuprind evaluarea prin comparare a două sau mai multe documente. Gradul de similitudine este folosit ca evaluare cantitativă al asemănării dintre două documente pe baza unui sistem de metrici. În lucrarea [2], este propusă o clasificare a principalelor metrici folosite în detecția plagiatului.

În literatura de specialitate, există două strategii principale de abordare a identificării plagiatului [3]:

- intrinsecă, care are scopul de a identifica pasaje plagiate prin examinarea numai a documentului supus analizei, concluzionând dacă părți ale materialului sunt sau nu scrise de acelaşi autor, astfel de modele fiind prezentate în [3] şi [4];

- externă, în care se presupune evaluarea prin comparare a documentului supus analizei cu celelalte documente de referință existente în baza de materiale şi identificarea perechilor de documente asemănătoare; multiple studii analizează această problematică, precum: [5], [6], [7].

Tehnica de identificare intrinsecă a plagiatului foloseşte stilul de scriere al unui autor ca bază de comparare. Se alcătuieşte un şablon format din caracteristici precum: statistici asupra textului, caracteristici sintactice, părți de vorbire, seturi de cuvinte folosite uzual sau caracteristici structurale ale textului. Setului de caracteristici i se ataşează o funcție criteriu de evaluare a modificărilor survenite de-a lungul textului analizat. Dezavantajele metodei sunt evidențiate în cazul lucrărilor scrise de mai mulți autori.

Pe de altă parte, abordarea externă a plagiatului aduce avantaje în sensul comparării documentului cu alte documente scrise de acelaşi autor, precum şi cu alte documente din acelaşi domeniu central. Dezavantajele sunt date de nivelul de complexitate exponențial în relație cu dimensiunea bazei de comparare.

2. Analiza stilometriei în plagiatul intrinsec

Identificarea intrinsecă a plagiatului presupune recunoaşterea acelor pasaje din cadrul unui text care, din punct de vedere al stilului de scriere, sunt diferite de celelalte pasaje. Aceste pasaje sunt ulterior analizate în scopul verificării de plagiat. Dacă un document este redactat de un singur autor, se presupune că toate pasajele redactate de el sunt similare în funcție de stilul unic de scriere.

Folosind aceeaşi tehnică de comparare a stilului de scriere din cadrul fiecărui pasaj al unei lucrări scrise de mai mulți autori, adăugând tehnici automate de clasificare nesupervizată de tip clusterizare, se pot grupa pasajele în funcție de autorii lucrării.

În cercetări precum cele elaborate în [8], [9], [5] şi [10] sunt tratate problemele şi metodele de integrare a plagiatului intrinsec făcând referire, de asemenea, la stilometrie, stilul de scriere al unui autor de-a lungul istoricului său de cercetare sau doar în cadrul unui document unitar.

Etapa de preprocesare a textului în scopul extragerii bogăției vocabularului constă în separarea în cuvinte a textului sau fragmentului de text analizat, eliminând spațiile, precum şi semnele de punctuație. O optimizare a prelucrării este dată şi de eliminarea cuvintelor de legătură, acestea fiind prezente în cadrul oricărui text redactat de diferiți autori. Notând cu W, mulțimea cuvintelor rezultate din această etapă de preprocesare, $W = \{w_1,w_2,\ldots,w_i,\ldots,w_N\}$, se introduce în analiză şi ontologia lexicală WordNet pentru generarea mulțimii de concepte unice regăsită din mulțimea inițială de cuvinte W în intersecție cu conceptele WordNet, prin reducerea dublurilor şi crearea vectorului de apariții a fiecărui concept regăsit, astfel:

$$\begin{cases} T = \{t_1, t_2, ..., t_i, ...t_{nt}\} \\ nap = \{nap_1, nap_2, ..., nap_i, ..., nap_{nt} \end{cases}$$

unde:

- T reprezintă mulțimea de concepte unice identificate din text şi regăsite în ontologia WordNet;
- nap reprezintă mulțimea formată din numărul de apariții al fiecărui concept din cadrul mulțimii T în cadrul documentului analizat.

În timp ce majoritatea indicatorilor de măsurare a bogăției vocabularului folosit de autorul unei lucrări se referă la relația dintre numărul de cuvinte unice identificate într-un text analizat raportat la numărul total de cuvinte existente în acel text, aceste metrici nu țin cont, în schimb, de componenta semantică existentă care derivă din acele cuvinte specifice extrase. De asemenea, metricile propuse şi extrase din literatura de specialitate nu evaluează evoluția în timp a acestei caracteristici care este transpusă într-un procent foarte mare în evaluarea stilului de scriere al unei persoane.

Pornind de la această problemă, este nevoie de propunerea unei metrici care să evalueze în acelaşi timp bogăția vocabularului folosit în cadrul respectivului document analizat, precum şi în cadrul documentelor anterioare, dacă acestea există. Metrica foloseşte numărul de cuvinte găsite, conceptele identificate folosind ontologia lexicală WordNet printr-o procesare de extragere a rădăcinii cuvintelor, precum şi funcțiile de calcul a distanțelor dintre oricare două concepte din WordNet.

Impactul folosirii acestei metrici este dat de latura semantică care este adăugată în cadrul mulțimii de cuvinte folosite într-un text analizat. Îmbogățind această metrică cu analiza semantică se generează o caracteristică de stilometrie complexă, din punct de vedere local, cât şi din punct de vedere al evoluției în timp.

Astfel, *IBSV*, Indicatorul de Bogăție Semantică al Vocabularului, este definit ca fiind egal cu:

$$IBSV = \frac{\sum_{i=1}^{nt} nap \times d\max(t_i)}{N}$$

unde:

- nap_i reprezintă numărul de apariții ale termenului unic aflat pe poziția i din mulțimea de termeni unici extraşi din documentul analizat;
- nt reprezintă cardinalitatea mulțimii de termeni unici extraşi din documentul analizat;
- $dmax(t_i)$ reprezintă distanța maximă dintre termenul unic t_i şi oricare alt termen unic extras din mulțimea de termeni, distanță care este calculată folosind distanțele semantice definite în cadrul ontologiei lexicale WordNet;
- N reprezintă cardinalitatea mulțimii cuvintelor, unice sau nu, extrase din documentul analizat rezultată în urma etapei de preprocesare a textului.

Indicatorul, $IBSV \in [0,1]$, iar o valoare a distanțelor $dmax(t_i) \to 0$, $\forall i = \overline{1,nt}$ conduce la o valoare a indicatorului $IBSV \to 0$. Interpretarea dată de acest context constă într-un document care este format din cuvinte, posibil distincte sau nu, dar care se regăsesc în aceeaşi arie semantică din punct de vedere al distanței semantice din cadrul ontologiei WordNet.

Situația opusă, cea în care $dmax(t_i) \to 1$, $\forall i = \overline{1,nt}$, transformă indicatorul propus în concordanță cu indicatoarele existente în literatura de specialitate folosite pentru măsurarea bogăției vocabularului utilizat de un autor în cadrul unui text sau fragment de text.

Tabelul 1 conține exemple de rulare a metricii propuse în scopul evaluării inițiale a rezultatelor obținute. Pentru evaluarea distanței dintre oricare două concepte din cadrul ontologiei WordNet este folosită metrica de tip Path Lenght, $d_{PATH}(c_1, c_2) = \frac{1}{\lg(c_1, c_2)}$, metrică care ia valori în intervalul [0,1]. De asemenea, este făcută o analiză comparată a metricii propuse cu metrica de tip Type – Token, metrică în format general acceptat de evaluare a bogăției vocabularului.

Tabel 1 - Rulare metrică IBSV pe un set de testare în comparație cu metrica de tip Type - Token

Fragment text analizat	**Vocabulary richness metrics are in depth analyzed in order to propose a new metric for evaluating the richness of the vocabulary used by authors of different documents by adding the semantic layer as a further characterization.**
Mulțime de cuvinte obținute în urma preprocesării	{Vocabulary, richness, metrics, are, in, depth, analyzed, in, order, to, propose, a, new, metric, for, evaluating, the, richness, of, the, vocabulary, used, by, authors, of, different, documents, by, adding, the, semantic, layer, as, a, further, characterization}
Mulțime de cuvinte rezultată în urma eliminării cuvintelor de legătură	{Vocabulary, richness, metrics, depth, analyzed, propose, new, metric, evaluating, richness, vocabulary, used, authors, different, documents, adding, semantic, layer, further, characterization}
Mulțime de concepte WordNet extrase	{Vocabulary, richness, metric, depth, analyze, propose, new, metric, evaluate, use, author, different, document, add, semantic, layer, further, characterization}

Rezultat metrică Type - Token	*Type – Token = 20/36 = 0.55*
Rezultat metrică *IBSV*	*IBSV = 0.26*
Analiză comparată	Metrica propusă, *IBSV*, ponderează rezultatul obținut de metrica *Type-Token* în sensul de similitudine semantică. Chiar dacă reducând cuvintele identificate în text la concepte WordNet unice, valoarea raportului este de *0.55 (55%),* nu este luată în considerare componenta de apropiere semantică. În textul analizat, există concepte din WordNet diferite sau apropiate ca similitudine, cu o valoare a distanței care tinde spre 0. Astfel, metrica *IBSV* exprimă în termeni mai realişti bogăția vocabularului regăsit în cadrul unui document sau fragment de text analizat.

3. Analiza evoluției în timp a stilometriei

Extinzând analiza dintre bogăția vocabularului şi distanța semantică dintre concepte cu evoluția în timp a acestei caracteristici orientate pe autor, se impune definirea trendului de evoluție al acestui indicator.

Contextul de analiză este dat de o mulțime inițială formată din documente redactate de un autor specific pentru care se face analiza, urmând a se înregistra valorile metricii propuse IBSV la nivelul fiecărui document. Această mulțime este sortată cronologic, în scopul generării unei serii de timp. Notând cu D, mulțimea inițială de documente analizate, unde $D=\{D_1,D_2,...,D_i,...D_n\}$ şi nd reprezintă cardinalitatea acestei mulțimi, se calculează mulțimea formată din valorile metricii *IBSV* aplicată la nivelul fiecărui document. Se menționează faptul că documentele din mulțimea D sunt sortate cronologic. Astfel, se obține seria de timp $IBSV=\{IBSV_1, IBSV_2,...,IBSV_i,...IBSV_{nd}\}$, având o cardinalitate egală cu mulțimea inițială de cuvinte.

Analiza îşi propune identificarea acelui trend pe care îl are valoarea indicatorului care măsoară bogăția semantică a vocabularului folosit de autor pe parcursul seriei de timp analizată. În situația în care există mai multe documente redactate de autor pe parcursul aceluiaşi an, valoarea indicatorului *IBSV* pentru anul respectiv se calculează ca medie aritmetică a valorilor indicatorului *IBSV* înregistrat în cadrul documentelor din ani egali.

Pentru a evalua trendul pe care indicatorul îl are la nivel de autor, se defineşte seria de timp folosind trei metode de estimare a trendului:

- *metoda modificării medii absolute* implică existența unei dependențe liniare de forma unei progresii aritmetice, în care fiecare termen al seriei este format pornind de la termenul inițial, termen prim din punct de vedere al timpului, prin adăugarea algebrică a unui decalaj multiplicat cu modificarea medie absolută; această metodă se pretează în contextul unei dependențe liniare de gradul întâi;

  $$\widehat{IBSV_{i1}} = IBSV_1 + \overline{\Delta} \times (t-1), \forall i = \overline{1,nd}$$

- *metoda indicelui mediu* implică existența unei dependențe exponențiale de forma unei progresii geometrice, în care fiecare termen al seriei este format pornind de la termenul inițial prin multiplicarea acestuia cu indicele mediu de dinamică exponențial; această metodă se pretează unei dependențe exponențiale între indicatorul *IBSV* şi seria perioadelor de timp;

  $$\widehat{IBSV_{i2}} = IBSV_1 \times (\overline{I})^{i-1}, \forall i = \overline{1,nd}$$

- *metoda modelului liniar de regresie* este singura metodă de tipul metodelor analitice propusă pentru analiză în exemplul de față şi implică estimarea unei ecuații de gradul întâi, estimare realizată pe baza metodei celor mai mici pătrate; forma trendului este dată de

formula:

$$\widehat{IBSV_{i2}} = \alpha + \beta \times i, \forall i = \overline{1, nd}$$

Alegerea celei mai bune metode de aproximare a trendului pe care îl are indicatorul *IBSV* de-a lungul timpului implică compararea sumei pătratelor erorilor generate de cele trei metode de estimare propuse:

$$\begin{cases} S_j = \sum_{i=1}^{nd} (IBSV_i - \widehat{IBSV_{ij}})^2, \forall j = \overline{1,3} \\ S^* = \max_{j=\overline{1,3}} S_j \end{cases}$$

Pentru detalierea analizei propuse, sunt extrase valorile indicatorului *IBSV* pentru un autor pe parcursul unei perioade de 13 ani, *nd=13*. Valorile sunt date de seria:

IBSV = {0.35; 0.37; 0.40; 0.29; 0.50; 0.45; 0.47; 0.39; 0.38; 0.47; 0.49; 0.50; 0,48}

Figura 1 reprezintă graficul evoluției indicatorului *IBSV* pe parcursul celor 13 ani.

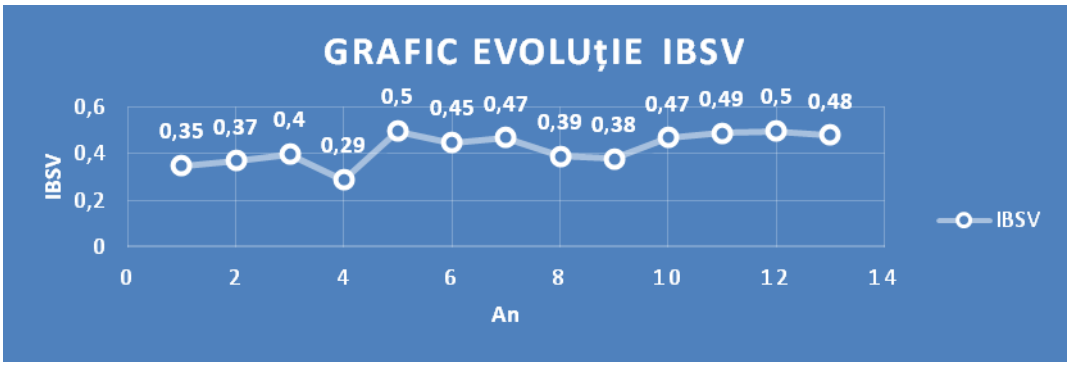

Figura 1. Graficul evoluției indicatorului *IBSV*

O analiză preliminară a graficului rezultat indică o evoluție crescătoare a valorii indicatorului *IBSV*, generând o interpretare la nivelul dezvoltării utilizării vocabularului printr-o creştere a bogăției sale la nivel semantic de analiză.

Pentru a putea previziona evoluția pe următoarea perioadă de cercetare, se impune rularea celor trei metode de extragere a trendului.

Sumarizând valorile obținute în analizele făcute, tabelul 2 conține suma pătratelor erorilor alături de ecuația obținută pentru estimarea trendului indicatorului *IBSV* pentru cele trei metode de estimare.

Tabel 2. Suma pătratelor erorilor şi ecuațiile trendului estimat pentru cele trei metode de estimare propuse

Metodă estimare	Suma pătratelor erorilor	Ecuație trend estimat
Metoda modificării medii absolute	$S_1 = 0.07$	$\widehat{IBSV_{i1}} = 0.35 + 0.01 \times (t-1)$
Metoda indicelui mediu	$S_2 = 0.15$	$\widehat{IBSV_{i2}} = 0.35 \times (1.02)^{i-1}$
Metoda regresiei liniare	$S_3 = 0.03$	$\widehat{IBSV_{i3}} = 0.34 + 0.011 \times i$

Cum pentru metoda folosind regresia liniară s-a obținut cea mai mică valoare a însumării pătratelor erorilor, se alege ca ecuație de estimare a trendului ecuația $\widehat{IBSV_{i3}} = 0.34 + 0.011 \times i$.

Figura 2 conține graficul trendului estimat în cadrul cercetării folosind regresia liniară. Se observă o evoluție crescătoare a trendului, cu 0.011 puncte procentuale de la o perioadă de timp la alta. Interpretarea este dată de o extindere a ariei de folosire a conceptelor extrase din cadrul documentelor redactate de autor.

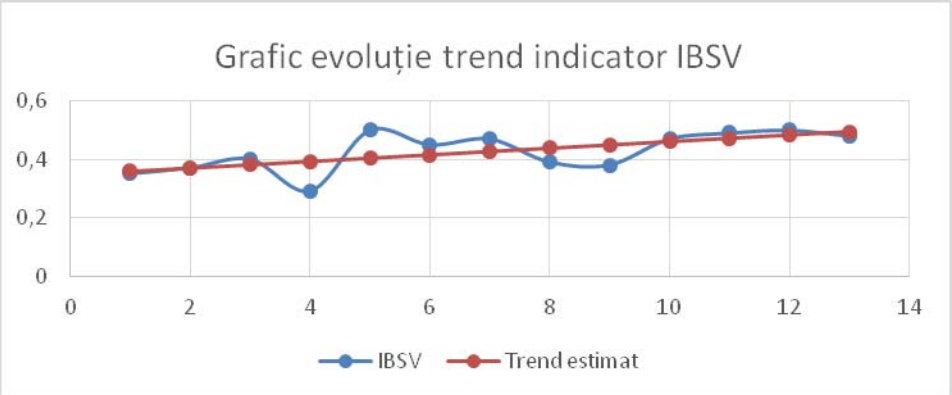

Figura 2. Graficul evoluției trendului indicatorului IBSV folosind regresia liniară

Pragul minimal al nivelului de originalitate în contextul analizei semantice este de 34%, observându-se o creştere liniară de 1.1% la nivelul evoluției de la un an la altul. Aproximarea trendului indică o direcție de cercetare extinsă de-a lungul perioadei de timp analizate, generând astfel o valoare cu un ritm crescător în cazul metricii propuse.

4. Concluzii

Metrica propusă, pornind de la metricile de tip Concept unic – Cuvinte şi ponderată cu distanțele semantice extrase pe baza ontologiei WordNet adaugă, pe lângă transformarea din cuvinte în concepte WordNet, şi distanțele semantice maximale dintre cuvinte, generând o componentă semantică neintrodusă în cercetările privind evaluarea, măsurarea şi interpretarea bogăției vocabularului folosit de un autor în cadrul unui text sau fragment de text redactat în limba engleză.

Avantajele aduse de prezenta metodă propusă constau în faptul că metrica de evaluare a bogăției vocabularului nu depinde de domeniile care sunt tratate în documentele analizate, ci doar de distanțele semantice dintre conceptele unice identificate în cadrul documentelor respective. Adăugând şi componenta timp în analiză, se întrevede o posibilă estimare a viitoarelor lucrări redactate de autori despre care se cunosc lucrări redactate anterior din punct de vedere al timpului.

* * *

Această lucrare a beneficiat de suport financiar prin proiectul "Rute de excelență academică în cercetarea doctorală şi post-doctorală – READ" cofinanțat din Fondul Social European, prin Programul Operațional Sectorial Dezvoltarea Resurselor Umane 2007-2013, contract nr. POSDRU/159/1.5/S/137926.

BIBLIOGRAFIE

1. **CEDENO, A. B.; VILA, M.; MARTI M. A.; ROSSO, P.:** Plagiarism Meets Paraphrasing: Insights for the Next Generation in Automatic Plagiarism Detection, Computational Linguistics, 39(4), pp. 917-947.
2. **LANCASTER, T.; CULWIN, F.:** Classifications of Plagiarism Detection Engines, Available at: http://www-new2.heacademy.ac.uk/assets/documents/subjects/ics/may2005_vol.4_1_classification_plagiarism_detection_engines.pdf

3. **OBERREUTER, G.; L'HUILLER, G.; RIOS, S. A.; VELASQUEZ J. D.:** Approaches for Intrinsic and External Plagiarism Detection, Notebook for PAN at CLEF, Available at: http://ceur-ws.org/Vol-1177/CLEF2011wn-PAN-OberreuterEt2011.pdf
4. **STAMATATOS, E.; KOPPEL, M.:**. Plagiarism and authorship analysis: introduction to the special issue, Lang Resources & Evaluation, 45(1), 2011, pp. 1-4.
5. **CARNAHAN, N.; HUDERLE, M.; JONES, N.; STEPHAN, C.; TRAN, T.; WOOD-DOUGHTY Z.:** Plagiarism Detection, 2014. Available at: http://www.cs.carleton.edu/cs_comps/1314/dlibenno/final-results/plagcomps.pdf
6. **ALZAHRANI, S. M.; SALIM, N.; ABRAHAM, A.:** Understanding Plagiarism Linguistic Patterns, Textual Features and Detection Methods, IEEE Transaction on Systems, Man and Cybernetics – Part C: Applications and Reviews, 42(2), 2012, pp. 133-149.
7. **SALUNKHE, S. D.; GAWALI, S. Z.:** A Plagiarism Detection Mechanism using Reinforcement Learning, International Journal of Advance Research in Computer Science and Management Studies, 1(6), 2013, pp. 125-129.
8. **EISSEN, S. M.; STEIN, B.; KULIG, M.:** Plagiarism Detection Without Reference Collections, Advances in Data Analysis Studies in Classification, Data Analysis and Knowledge Organization, 2007, pp. 359-366.
9. **STEIN, B.; LIPKA, N.; PRETTENHOFER, P.:** Intrinsic plagiarism analysis, Language Resources & Evaluation, 45(1), 2010, pp. 63-82.
10. **SEDDING, J.; KAZAKOV, D.:** WordNet-based Text Document Clustering, ROMAND 2004, Workshop on Robust Methods in Analysis of Natural Language Data, Geneva, August, 2004, pp. 104-113.
11. **OSICEANU, M-E.:** Consideraţii privind drepturile de proprietate intelectuală în ştiinţă, tehnică şi artă sau între creaţie şi plagiat, Available at: http://api.ning.com/files/uPa7BpseSwF6lqvQmgiaPdijUqzZEL9nHLQzkOJht94wzdjkfubWxs5cGMbkITg3agVjj0s2dOhxhjn88Hy*72*M4OH2MIVb/Osiceanu_MEConsideraiiprivinddrepturiledeproprietateintelectuala_final.pdf