METRICA REPREZENTATIVITĂȚII ÎN ENTITĂȚI TEXT

Ion Ivan ionivan@ase.ro
Daniel Milodin daniel.milodin@ase.ro
Mihai Georgescu mihaita.georgescu@gmail.com
Academia de Studii Economice Bucureşti

Abstract: Se definesc conceptele de entităţi text structurate şi componentele acestora. Se detaliază conceptul de ortogonalitate a entităţilor structurate şi se aplică acest concept în cadrul textelor componente ale aceleaşi entităţi. Sunt definite criteriile ce stau la baza studiului ortogonalităţii. Se detaliază operaţiunea de normalizare a textelor, în vederea eficientizării analizelor efectuate. Se prezintă conceptul de repetitivitate a subşirurilor text**.**

Cuvinte cheie: entităţi structurate, ortogonalitate, repetitivitate.

Abstract: Are defined the concepts of structured text entities and their components. Is presented in detail the orthogonality concept of structured entities and it is applied in the text components of the same entity. Are defined the criteria underlying the study of orthogonality. Is detailed the operation of normalizing the texts in order to make the performed analysis more efficient. Is presented the concept of text substrings repetition.

Keywords: structured entities, orthogonality, repetition.

1. Entităţi text structurate

Entităţile structurate creează şi implementează forme de stocare şi prelucrare care sunt folosite pentru gestiunea informaţiei.

Organizarea informaţiei se realizează în funcţie de caracteristicile care o definesc şi care o individualizează.

Structurarea entităţilor reflectă asocierile existente între date.

Entităţile structurate sintetizează modalitatea de grupare a caracteristicilor, precum şi a legăturilor dintre acestea.

Între elementele componente ale unei entităţi structurate există o relaţie de dependenţă.

Fiecare componentă a entităţii are un corespondent în plan real, materializat prin niveluri sau subniveluri de organizare.

Structurarea entităţilor este realizată pe baza dispunerii elementelor ce compun sistemul analizat, structurile construite lucrând cu informaţii care sunt stocate în cadrul arborescenţei implementate. Condiţiile ce trebuie respectate în cadrul structurării implică dispunerea entităţilor pe nivele şi prelucrarea informaţiilor stocate pe fiecare nivel.

Organizarea este realizată pe diferite domenii de activitate, definirea entităţilor text făcându-se în funcţie de criteriul de structurare şi de informaţiile cu care operează.

Entităţile text structurate sunt construite pe baza unor concepte de bază, precum [5]:

- simbolul constituie o modalitate de reprezentare; prin folosirea simbolurilor se pun în corespondenţă obiecte, noţiuni, imagini cu reprezentarea acestora; simbolul sugerează elemente reale prin înlocuirea acestora cu anumite forme considerate reprezentative; simbolul este definit printr-un singur semn, α reprezentând litera grecească alfa, sau printr-un grup de semne, H_2O reprezentând simbolul chimic al apei; cu ajutorul simbolului sunt reprezentate cantităţi, fenomene, imagini, idei, operaţii, formule; de-a lungul timpului modalitatea de utilizare a simbolului a evoluat, în funcţie de gradul de dezvoltare a societăţii umane; iniţial, simbolul a definit obiecte, oameni, animale, rolul lui fiind acela de a constitui elemente ale unui alfabet în procesul de comunicare; treptat, o dată cu apariţia literelor şi alfabetelor, simbolul şi-a pierdut din importanţă, acesta trecând în plan secundar;
- alfabetul este o mulţime finită, alcătuită din literele folosite în scrierea unei limbi; caracteristica de bază a literelor ce redau cuvintele unei limbi este de a respecta o ordine

convenţională; funcţie de simbolurile utilizate alfabetele se clasifică în alfabete fonetice şi alfabete simbolice precum alfabetul Morse, alfabetul Braille; plecând de la rolul de bază al alfabetului, acela de a reda sunetele unei limbi, implicaţiile alfabetului s-au extins asupra unor arii noi de interes; în informatică alfabetul defineşte totalitatea simbolurilor care stau la baza unui limbaj de programare;

- cuvântul reprezintă asocierea dintre un sens şi un complex sonor; cuvântul este unitatea de bază a vocabularului; în funcţie de modalitatea de formare cuvintele se clasifică în: cuvinte de bază – cuvinte care servesc ca elemente de bază în formarea altor cuvinte, cuvinte compuse – cuvinte formate prin compunerea mai multor cuvinte şi cuvinte derivate – cuvinte formate plecând de la un alt cuvânt; unui cuvânt îi corespund unul sau mai multe sensuri;
- vocabularul este format din mulţimea cuvintelor specifice unei limbi; frecvenţa de utilizare a cuvintelor împarte vocabularul în:
  - vocabular activ alcătuit din totalitatea cuvintelor specifice unei limbi folosite în mod efectiv în exprimare;
  - vocabular pasiv ce conţine cuvintele specifice unei limbi care nu sunt utilizate în mod frecvent;
  - vocabular fundamental alcătuit din fondul principal de cuvinte folosite;
  - vocabular secundar ce formează masa vocabularului; de asemenea, termenul de vocabular defineşte cuvintele specifice categoriilor sociale, domeniilor de activitate, de cercetare;
- separatorul este un simbol care nu aparţine alfabetului, cu rolul de a delimita cuvintele din alfabet când sunt folosite pentru transmiterea de informaţii; textele sunt construite prin utilizarea de separatori; separatorii : ; ? ! @ # $ % ^ & * ( ) ` ~ , . < > / \ - _ + = | " ' [ ] " " } { sunt alocaţi mulţimii separatorilor;
- frecvenţa de apariţie este un indicator care furnizează informaţii despre nivelul de utilizare a termenilor din vocabularul de bază în cadrul entităţilor structurate; în funcţie de numărul de apariţii, termenii se împart în:
  - termeni cu frecvenţă scăzută sunt termenii care la un anumit număr de cuvinte au un număr scăzut de apariţii;
  - termeni cu frecvenţă ridicată de apariţie sunt termenii care la un număr dat de cuvinte au un număr ridicat de apariţii;
- textul este alcătuit din mai multe cuvinte separate care transmit o anumită informaţie; textul este caracterizat de lungime, măsurată în număr de cuvinte sau în număr de simboluri folosite, de frecvenţa de apariţie a cuvintelor sau simbolurilor, precum şi de gradul de apartenenţă la un vocabular prin care se stabileşte în ce măsură textul respectiv foloseşte termeni dintr-un domeniu; textul *Ortogonalitatea textelor oferă informaţii despre nivelul de originalitate al textelor şi despre nivelul de asemănare al textelor* conţine un număr de 17 cuvinte, din care 11 cuvinte distincte şi 112 caractere;
- tezaurul este mulţimea formată din totalitatea cuvintelor considerate definitorii pentru un domeniu;
- sistemul referit defineşte ansamblul de elemente structurate, studiat şi implementat folosind entităţile structurate.

Gradul de dificultate a conceptelor folosite în cadrul entităţilor oferă indicii cu privire la nivelul de reprezentare folosit. Cuvântul, simbolul şi separatorii sunt folosiţi în cuprinsul textelor şi în construirea de entităţi text. Vocabularul şi tezaurul sunt create ţinând seama de apartenenţa cuvintelor folosite la o arie de interes definită cu ajutorul tezaurului şi a vocabularului [1].

Modelele de stocare a textelor construite în mod flexibil şi cu deschidere pentru elementele de noutate permit prelucrarea de diverse tipuri de texte, orientate spre diverse domenii de activitate.

Entităţile text structurate sunt în general folosite pentru lucrul cu proiectele.

Un text este structurat când este alcătuit din mai multe subtexte ce au o existenţă de sine stătătoare.

Şirurile de cuvinte sunt caracterizate prin poziţii ale cuvintelor, prin legăturile existente între acestea, prin definirea de contexte şi prin existenţa unei corespondenţe între aceste cuvinte şi lumea reală.

Entităţile text au o arie largă de aplicabilitate, pornind de la articole simple, de ziar, până la produse program scrise în diverse limbaje.

Dimensiunea unui vocabular este dată de numărul cuvintelor componente. Vocabularul limbii române este compus dintr-un număr relativ mic de cuvinte, aproximativ 200.000 de cuvinte, comparativ cu limba engleză care a depăşit recent un număr de 1.000.000 cuvinte.

DEX - Dicţionarul explicativ al limbii române 2009 conţine un număr de 65000 de cuvinte (cele mai utilizate în limba română contemporană şi în operele literare clasice), precum şi un mare număr de regionalisme, arhaisme şi neologisme, sensuri şi unităţi frazeologice noi.

Fiecare cuvânt al unei limbi reprezintă o legătură între un obiect, o stare, o conjunctură şi un limbaj universal acceptat de vorbitorii acelei limbi. Prin intermediul cuvântului sunt create premisele unei interfeţe care asigură comunicarea între membri societăţii.

Astfel, cu cât vocabularul unei limbi este mai variat, cu atât comunicarea se realizează mai eficient, mai exact.

2. Ortogonalitatea internă a entităţilor structurate

Ortogonalitatea studiază gradul de asemănare între două sau mai multe entităţi. Prin intermediul acestei caracteristici de calitate este determinată măsura în care entităţile diferă una de cealaltă [2].

Conceptul de ortogonalitate provine din domeniul matematicii, unde are în vedere aspectele următoare:

- două plane sunt ortogonale dacă unghiul format la intersecţia lor are cosinusul egal cu valoarea zero; o mulţime finită de plane este ortogonală dacă planele sunt perpendiculare două câte două;
- două drepte sunt ortogonale dacă formează unghiuri adiacente congruente; o mulţime finită de drepte este ortogonală dacă dreptele sunt perpendiculare două câte două;
- doi vectori sunt ortogonali dacă produsul scalar al acestora este nul; o mulţime finită de vectori este ortogonală dacă produsul scalar al oricăror doi vectori diferiţi este nul.

Ortogonalitatea este studiată pe baza criteriilor de ortogonalitate. Cu ajutorul acestor criterii sunt evidenţiate caracteristicile care au aceeaşi valoare pentru entităţile studiate şi sunt determinate nivelurile de asemănare.

Ortogonalitatea entităţilor text este studiată ţinând cont de:

- conţinutul lor informaţional;
- semnele utilizate pentru a reda informaţia, respectiv de semiotica entităţilor.

Pentru a fi studiată ortogonalitatea a două sau mai multe entităţi, ca o primă condiţie ce trebuie respectată, acestea trebuie să aibă aceeaşi structură, adică să fie definite prin intermediul aceloraşi caracteristici.

Plecând de la caracteristicile entităţilor, sunt definiţi indicatori de calitate ai acestora. Pe baza acestor indicatori este construit un indicator ce ţine cont de valorile indicatorilor componenţi.

Un domeniu în care ortogonalitatea este foarte importantă este programarea. Limbajele de programare sunt astfel proiectate încât ortogonalitatea acestora să fie maximă, în sensul că noţiunile implementate, termenii şi cuvintele cheie folosite sunt unice pentru a nu crea confuzie, atât în rândul utilizatorilor, cât mai ales pentru a nu aduce compilatoarele în situaţia de a nu cunoaşte ce presupune o linie cod, din cauza mai multor opţiuni posibile [3].

Compararea a două entităţi se reduce la a raporta o entitate la cealaltă entitate, respectiv la a identifica părţile comune şi părţile care diferă. Sunt astfel, comparate caracteristicile

corespondente ale celor două entități.

Pentru entitățile text un rol important îl are frecvența de apariție a cuvintelor în cadrul entităților. Prin intermediul frecvenței este determinată importanța cuvintelor în cadrul entităților şi este determinat gradul de utilizare a cuvintelor, precum şi felul cum acestea influențează construirea entităților.

Ortogonalitatea textelor se determină ca ortogonalitate internă şi ortogonalitate externă. Ortogonalitatea internă stabileşte măsura în care, în cadrul textului, cuvintele folosite se repetă, identificând legăturile existente între cuvintele din cadrul textului, precum şi modalitatea de dispunere a acestora.

Ortogonalitatea externă arată care sunt diferențele sau asemănările ce există între texte.

Ortogonalitatea entităților text este studiată pe baza următoarelor criterii [4]:

- dimensiunea fişierelor care stochează entitățile text, exprimată ca număr de biți;
- frecvențele de apariție a cuvintelor în cadrul textelor, calculate prin identificarea numărului de apariții ale cuvintelor în cadrul textelor;
- frecvențele de apariție a cuvintelor de legătură, calculate prin identificarea numărului de apariții ale cuvintelor de legătură în cadrul textelor;
- dimensiunea textelor privită ca număr de cuvinte sau ca număr de litere care alcătuiesc textele;
- dispunerea cuvintelor în cadrul textelor, respectiv stabilirea poziției fiecărui cuvânt şi identificarea similitudinilor;
- distanțele dintre cuvinte definite ca număr de cuvinte care se interpun între două cuvinte considerate;
- numărul de paragrafe ale textelor.

Două entități text sunt ortogonale dacă nu au cuvinte comune.

Ortogonalitatea a două texte este calculată prin inventarierea cuvintelor comune.

În cazul studierii ortogonalității foarte important este şi sensul cu care sunt folosite datele. Dacă sunt folosite entități pentru a defini un domeniu de activitate şi în cadrul entităților sunt anumite restricții privitoare la datele ce sunt disponibile, atunci frecvențele de utilizare vor avea valori apropiate, conținutul va fi asemănător, singurul aspect care deosebeşte entitățile fiind sensul cu care sunt folosite datele.

Este urmărită modalitatea în care două entități diferă sau sunt asemănătoare, atât ca frecvențe de apariție a cuvintelor, conținut, precum şi sens al datelor conținute.

Conceptul de ortogonalitate se aplică pentru a identifica dacă anumite lucrări aparțin sau nu unui domeniu de referință. Este studiat gradul de asemănare, frecvențele de utilizare a cuvintelor, iar pe baza indicatorului rezultat se determină dacă textele sunt asemănătoare sau nu. În situația în care indicatorul de ortogonalitate tinde spre 1, adică lucrările sunt total diferite, este evident că nu aparțin aceluiaşi domeniu. Pentru a identifica în mod corect dacă două sau mai multe lucrări aparțin aceluiaşi domeniu trebuie identificate cuvintele de specialitate conținute de lucrări şi frecvențele de apariție a acestora.

Comparabilitatea textelor se analizează la următoarele niveluri:

- formă de reprezentare;
- conținut informațional.

Reprezentarea percepțiilor omului privind lumea înconjurătoare se realizează prin intermediul construcțiilor sintactice formate din simboluri alfanumerice, grafice, spațiale.

Criteriile de comparare a două entități vizează:

- lungimea entității text – este exprimată în mai multe moduri:
    - număr de simboluri utilizate;

    - număr de caractere utilizate cu posibilitatea de a diferenția pe caractere mici şi caractere mari;
    - număr de cuvinte folosite;
    - număr de pagini;
    - număr de bytes ocupați, în cazul entităților text care sunt date în format electronic.
- lungimea vocabularelor entităților – exprimată ca număr de cuvinte distincte utilizate în construcția entităților text;
- lungimea vocabularelor cuvintelor cheie – sunt luate în considerare cuvintele definitorii pentru domeniul abordat;
- numărul de cuvinte cheie din titlurile entităților;
- diversitatea vocabularelor entităților – raportează cuvintele din vocabularele entităților cu cele din vocabularul limbii în care au fost realizate;
- frecvențele de apariție a cuvintelor din entități – pentru fiecare cuvânt din vocabularul unei entități se asociază numărul de apariții în cadrul entității;
- frecvențele de apariție a cuvintelor din vocabularele entităților în raport cu un vocabular definit de utilizator;
- construirea matricelor de precedențe în vederea analizei comparate a textelor în funcție de pozițiile construcțiilor sintactice.

Comparabilitatea textelor din punctul de vedere al conținutului informațional are în vedere următoarele aspecte:

- delimitarea domeniului abordat prin evidențierea semanticii cuvintelor cheie;
- înțelesul cuvintelor specifice domeniului şi sensul în care acestea au fost utilizate în construcția entității text;
- acoperirea cadrului conceptual prin utilizarea cuvintelor specifice domeniului;
- aducerea entităților la o formă comună prin eliminarea cuvintelor de legătură şi a prefixelor/sufixelor adăugate conform regulilor gramaticale specifice limbii;
- identificarea sinonimelor pentru cuvintele din vocabularul entităților şi modificarea textelor prin utilizarea aceleiaşi forme a cuvântului;
- identificarea şi prezentarea elementelor de natură teoretică şi practică exprimate în cele două entități şi realizarea de asocieri;
- identificarea funcționalităților descrise în texte, inventarierea lor şi evidențierea eventualelor asocieri.

Ortogonalitatea entităților text este studiată prin prisma domeniului căruia îi aparțin textele prelucrate cu ajutorul entităților.

Două entități text sunt ortogonale dacă nu au cuvinte comune.

Versurile următoare nu conțin elemente repetitive:

La steaua care-a răsărit
E-o cale-atât de lungă,
Că mii de ani i-au trebuit
Luminii să ne-ajungă.

Spre deosebire de următoarele versuri în care se repetă primul şi ultimul vers:

Vreme trece, vreme vine,
Toate-s vechi şi nouă toate;
Ce e rău şi ce e bine
Tu te-ntreabă şi socoate;
.....
Tu aşează-te deoparte,
Regăsindu-te pe tine,
Când cu zgomote deşarte
Vreme trece, vreme vine.

Ortogonalitatea a două texte este calculată prin inventarierea cuvintelor comune. Procedeul devine complex când textele aparţin aceluiaşi domeniu sau au valoare ştiinţifică.

Compararea a două texte ştiinţifice pentru stabilirea ortogonalităţii implică determinarea schemei de dispunere a cuvintelor. Trebuie identificată modalitatea de folosire a cuvintelor şi de poziţionare a acestora unele faţă de altele.

Fiecare autor are un stil propriu de îmbinare a cuvintelor ce aparţin domeniului ştiinţific.

Ortogonalitatea entităţilor text implică aspecte precum:

- determinarea poziţiei cuvintelor;
- determinarea distanţelor dintre cuvinte;
- stabilirea frecvenţei cuvintelor care aparţin vocabularului domeniului.

La studierea ortogonalităţii trebuie să se ţină cont de contribuţia pe care autorul o aduce în text.

Pentru o analiză eficientă, textele trebuie normalizate.

Normalizarea are în vedere următoarele operaţii:

- eliminarea secvenţelor de separatori şi a caracterelor de control – presupune ştergerea secvenţelor redundante de simboluri care marchează zone de text;
- eliminarea diferenţierii între caracterele mari şi caracterele mici pentru uniformizarea comparaţiilor;
- eliminarea folosirii diacriticelor – constă în înlocuirea caracterelor specifice unei limbi cu caractere echivalente;
- înlocuirea simbolurilor din alte limbi – presupune stabilirea corespondenţei între cuvintele ce nu aparţin limbii şi succesiunea de caractere specifice acesteia.

Prin normalizare este asigurată omogenitatea datelor din punct de vedere al comparabilităţii lor.

Ortogonalitatea entităţilor text este studiată din punct de vedere semantic şi din punct de vedere semiotic.

Ortogonalitatea semantică studiază cuvintele ca şi sens al acestora, respectiv stabileşte dacă două sau mai multe cuvinte diferă ca înţeles. Pentru aceasta este determinată rădăcina cuvintelor şi se stabileşte modul de evoluţie al fiecărui cuvânt, pornind de la rădăcină.

Ortogonalitatea semiotică determină în ce măsură textele se aseamănă din punct de vedere al simbolurilor utilizate, al poziţiei în cadrul textului şi a distanţei dintre cuvinte în cadrul textului.

3. Repetitivitatea subşirurilor identice în cadrul entităţilor text

Conceptul de repetitivitate defineşte calitatea entităţilor de a fi alcătuite din texte care apar în cadrul acestora la poziţii diferite, dar având aceeaşi structură a dispunerii.

Se consideră entitatea text E_1 alcătuită din textele T_1, T_2 şi T_3, astfel:

T_1 = *{aaa, bbbe, ccc}*

T_2 = *{, ppap, ss, mmm, bbbe, ss, nnnc, ss, pprp, nn, ddd, ccc, bbbe, aaa, ii }*

T_3 = *{ ppap, mmm, bbbe, ss, nnnc, ss, pprp},*

entitatea E_1 fiind formată din reuniunea celor trei texte, astfel:

$E_1 = T_1 + T_2 + T_3$ = *{aaa, bbbe, ccc, ii, ppap, ss, mmm, bbbe, ss, nnnc, ss, pprp, nn, ddd, ccc, bbbe, aaa, ii, ppap, mmm, bbbe, ss, nnnc, ss, pprp },*

unde operaţia de reuniune este reprezentată prin convenţie de simbolul +.

Lungimea entităţii E_1 este dată de numărul total al cuvintelor componente, respectiv 25.

Textele T_1, T_2 şi T_3 sunt construite pe baza cuvintelor:

$<c_1>$ = { *aaa* }

$<c_2>$ = { *bbbe* }

$<c_3>$ = { *ccc* }

$<c_4>$ = { *ppap* }

$<c_5>$ = { *mmm* }

$<c_6>$ = { *nnnc* }

$<c_7>$ = { *pprp* }

$<c_8>$ = { *nn* }.

Pornind de la cuvintele identificate în mod unic în cadrul celor trei texte componente ale entității, se construieşte vocabularul reprezentativ al entității, VE, alcătuit din cele 8 cuvinte, astfel:

VE = { *aaa, bbbe*, *ccc*, *ppap*, *mmm*, *nnnc*, *pprp*, *nn* }.

De asemenea, cele trei texte conțin şi cuvinte cu o frecvență ridicată de apariție, respectiv cuvinte de legătură sau cuvinte cu caracter comun. Aceste cuvinte fac parte din vocabularul VL al cuvintelor de legătură:

VL = { *ii, ss, ddd,*}.

În vederea unei analize eficiente se recomandă implementarea unor algoritmi de normalizare cu rolul de a înlătura atât cuvintele de legătură, cât şi cuvintele care au o frecvență ridicată de apariție ce afectează analiza.

După normalizare, entitatea E_1 se prezintă astfel:

E_1 = {*aaa, bbbe, ccc, ppap, mmm, bbbe, nnnc, pprp, nn, ccc, bbbe, aaa, ppap*, *mmm, bbbe, nnnc, pprp* },

noua lungime fiind de 17 cuvinte.

În vederea stabilirii subşirurilor de cuvinte care se repetă în cadrul entității E_1 se propune următorul algoritm:

P_1: se construieşte vocabularul ataşat entității text, în cazul prezentat vocabularul VE;

P_2: se construieşte vocabularul pentru cuvintele de legătură şi pentru cuvintele cu frecvență ridicată de apariție, în exemplul considerat VL;

P_3: se normalizează entitatea E_1;

P_4: pornind de la vocabularul VE se parcurge textul entității E_1 şi se notează poziția fiecărui cuvânt al entității în cadrul vocabularului, rezultând matricea asociată M;

P_5: pornind de la prima coloană se parcurge matricea, iar unde se identifică valoarea 1 în cadrul coloanei parcurse, semn că un cuvânt din vocabular este prezent în cadrul textului, se trece la identificarea următoarei coloane care stochează valoarea 1; se construieşte în acest fel un subşir care este căutat în matricea M coloană după coloană, pornind de la linia unde a fost identificat primul cuvânt al subşirului; la identificarea de poziții consecutive identice se incrementează contorul asociat subşirului pentru a i se stabili lungimea maximă; căutarea subşirurilor se realizează prin incrementare şi prin decrementare, cuvintele formând acelaşi subşir şi în ordine inversă ca în exemplul: T_1 = { *annn arrr, ammm* } şi T_2 = { *ammm, arrr, annn*}

P_6: se calculează gradul de repetitivitate al subşirurilor;

Conceptul de repetitivitate este o particularizare a conceptului de ortogonalitate în cadrul aceluiaşi text, prin intermediul acestuia identificându-se gradul de originalitate internă a unui text şi asigurându-se construirea unor înlănțuiri de cuvinte diferite.

Pornind de la entitatea E_1 se construieşte matricea M asociată identificării cuvintelor vocabularului VE în cadrul acestei entități:

Tabelul nr. 1 Matricea M asociată pozițiilor cuvintelor vocabularului VE în cadrul entității E_1

Poziția în text / Cuvinte vocabular	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
aaa	1											1					
bbbe		1				1					1				1		
ccc			1							1							
ppap				1									1				
mmm					1									1			
nnnc							1									1	
pprp								1									1
nn									1								

Având la bază algoritmul descris anterior se porneşte cu prima coloană a matricei, întrucât pe prima coloană se regăseşte primul cuvânt al vocabularului VE. Se observă că primul cuvânt, respectiv *aaa* se regăseşte în cadrul textului pe poziţiile 1 şi 11. Aceste valori se identifică urmărind poziţiile pe care se află stocată cifra 1 în cadrul liniei corespondente a cuvântului analizat. Pornind de la aceste două coloane, se incrementează sau decrementează valorile poziţiilor unde a fost regăsit cuvântul *aaa*, această operaţiune efectuându-se atât timp cât în cadrul matricei M valorile identificate pe poziţiile incrementate / decrementate sunt identice.

Rezultă, astfel că, pornind de la cuvântul *aaa* se construieşte subşirul *aaa, bbbe, ccc*, respectiv subşirul în formă inversă *ccc, bbbe, aaa*, precum şi subşirul *ppap, mmm, bbbe, nnnc, pprp*, regăsit în forma unui subşir identic în conţinutul entităţii E_1.

De asemenea, mai sunt identificate şi alte subşiruri, de lungime mai mică decât cele anterioare, acestea fiind incluse în primele două subşiruri.

Modul de lucru al algoritmului prezentat se bazează pe caracterul unic al apariţiei unui cuvânt al vocabularului pe o anumită poziţie din cadrul textelor ce compun entitatea structurată. Astfel, pe aceeaşi poziţie într-o text va fi identificat un singur cuvânt. Implementarea algoritmului asigură identificarea subşirurilor de cuvinte care se repetă în cadrul unui text, pe baza stabilirii succesiunii de apariţii ale cuvintelor. Se identifică în acest mod o caracteristică a modului de utilizare a cuvintelor specifice fiecărui text sau se stabileşte dacă un text este format prin utilizarea în mod repetat a aceloraşi subşiruri.

În cadrul entităţilor structurate de tip articol, repetitivitatea ia în considerare aspecte care asigură definirea şi implementarea noţiunilor prezentate în partea introductivă a textului:

- cuvintele cheie se regăsesc în cadrul articolului;
- unele cuvinte cheie se regăsesc în titlurile articolelor din bibliografie;
- cuvintele cheie se regăsesc în subtitlurile articolului.

Aplicabilitatea directă a conceptului de repetitivitate în cazul entităţilor structurate de tip articol este dată de următoarele reguli:

- în titlul unui articol cuvintele nu se repetă;
- în cadrul secţiunii *abstract* o frază apare o singură dată;
- în capitolele componente ale articolului o definiţie este oferită o singură dată.

Implementarea repetitivităţii asigură creşterea calităţii entităţilor structurate prin definirea de criterii de calitate rezultate din studierea modalităţii de dispunere a cuvintelor în cadrul textelor. Criteriile de calitate reflectă nivelul de repetare al cuvintelor în cadrul textelor, modalităţile de dispunere a cuvintelor în componenţa paragrafelor, şi gradul de apartenenţă a cuvintelor la un vocabular de specialitate.

4. Indicatori ai repetitivității

În vederea stabilirii nivelului de ortogonalitate, având la bază algoritmul descris mai sus, se definesc indicatori care certifică originalitatea unui text.

Astfel, se defineşte un indicator normat de ortogonalitate cuprins în intervalul [0, 1], care ia următoarele valori:

- 1, dacă elementele sunt ortogonale, adică nu au nimic în comun;

- 0, elementele sunt identice, adică nu au valori diferite pentru nicio caracteristică.

Dacă valoarea indicatorului tinde către 1 înseamnă că seturile de date conținute de cele două entități tind către ortogonalitate, iar dacă valoarea indicatorului este apropiată de 0 înseamnă că seturile de date au foarte multe elemente identice.

Problema ortogonalității textelor este luată în considerare pentru validarea textelor. În vederea analizării textelor se implementează reguli ce trebuie respectate pentru construirea textelor. Astfel, textele componente ale entității sunt delimitate de separatori şi se încheie cu semnul grafic ".". Entitatea structurată este compusă din mai multe texte, care se îmbină pentru a evidenția o informație.

Se defineşte indicatorul de repetitivitate a subşirurilor componente ale unei entități structurate:

$$H(\text{subşir}_A,\text{subşir}_B) = \frac{LC}{\max(L_A, L_B)},$$

unde:

LC – numărul cuvintelor comune celor două subşiruri;
L_A – lungime subşir_A;
L_B – lungime subşir_B.

Determinarea repetitivității este raportată la dimensiunile textelor care se analizează şi nu la dimensiunea totală a textului reunit, fapt care conduce către un rezultat cu un grad de reprezentativitate mai ridicat.

Se consideră cele trei texte, normalizate:

T_1 = {*aaa, bbbe, ccc*}

T_2 = { *ppap, mmm, bbbe, nnnc, pprp, nn, ccc, bbbe, aaa* }

T_3 = { *ppap, mmm, bbbe, nnnc, pprp*},

precum şi cele două subşiruri comune identificate în cadrul celor trei texte:

S_1 = {*aaa, bbbe, ccc*}

S_2 = { *ppap, mmm, bbbe, nnnc, pprp* },

cele două subşiruri regăsindu-se în cadrul textelor T_1 şi T_2, pentru subşirul S_1, respectiv T_2 şi T_3, pentru subşirul S_2.

Dimensiunile celor trei texte normalizate sunt următoarele:

$L(T_1) = 3$;

$L(T_2) = 9$;

$L(T_3) = 5$,

$L(S_1) = 3$;

$L(S_2) = 5$, cele două subşiruri reprezentând cuvintele comune subşirurilor.

$H(T_1,T_2) = 3 / 9 = 0.33$,

$H(T_2,T_3) = 5 / 9 = 0.56$.

În vederea obţinerii unei valori care să cumuleze valorile parţiale ale ortogonalităţilor, obţinute prin compararea textelor luate două câte două, se consideră indicatorul agregat, construit pe baza formulei:

$$HG = 1 - \left(\prod_{i=1}^{NS-1} \prod_{j=i+1}^{NS} HG(sir_i, sir_j) \right)^{\frac{1}{NS(NS-1)}}$$

unde:

NS – numărul de şiruri care compun entitatea text analizată;

$Şir_i$ – şirul *i* care este comparat cu un alt şir component al entităţii.

Pentru exemplul considerat valoarea indicatorului agregat HG este 0.24, ceea ce arată că nivelul ortogonalităţii interne a entităţii E_1 este foarte mic, şirurile fiind utilizate în mod excesiv la construirea entităţii.

Repetitivitatea subşirurilor este urmărită pentru secvenţe de texte complete, precum propoziţiile şi frazele, nu doar pentru şiruri alcătuite din foarte puţine cuvinte. Se urmăreşte astfel identificarea repetitivităţii care influenţează originalitatea textelor.

5. Concluzii

Conceptul de repetitivitate identifică gradul de reutilizare al subşirurilor în cadrul unui text. Sunt identificate atât caracterul original al textelor, cât mai ales tendinţele de a refolosi subşirurile în vederea construirii unui text. Algoritmul prezentat asigură determinarea gradului de reutilizare al textelor pe baza vocabularului construit pornind de la cuvintele ce compun textul analizat. Se asigură astfel eficientizarea stabilirii ortogonalităţii interne a textelor.

În vederea obţinerii de rezultate cu o acurateţe ridicată sunt identificate şi înlăturate cuvintele de legătură, precum şi cuvintele cu grad ridicat de repetitivitate, care influenţează indicatorii rezultaţi.

Studierea conceptului de repetitivitate în cadrul textelor asigură creşterea gradul de diversitate şi implicit creşterea gradului de originalitate, prin utilizarea de texte cu caracter unic, orientate spre aceeaşi problematică.

BIBLIOGRAFIE

1. **IVAN, I., M. POPA**: Text Entities – Development, Evaluation, Analysis, ASE Printing House, Bucharest, 2005.
2. **POPA, M.:** Evaluarea calităţii entităţilor text, Teorie şi Practică, Editura ASE, Bucureşti, 2005.
3. **SMEUREANU, I., M. DARDALA**: Programarea în limbajul *C/C++*, Editura CISON, Bucureşti, 2004.
4. **IVAN, I., D. MILODIN, M. POPA**: Operation on text entities, Informatica Economica, vol. 11, nr. 1, 2007, pp. 14 – 20.
5. **MILODIN, D.**: – Ortogonalitatea entităţilor structurate – Teză de doctorat, noiembrie 2009.