ENTITĂŢI TEXT ŞI CARACTERISTICI DE CALITATE. DEFINIREA CONTEXTULUI, A ENTITĂŢILOR ORIENTATE PE TEXT, A CARACTERISTICILOR DE CALITATE

Prof. dr. Ion Ivan, Asist. univ. Marius Popa, Asist. univ. Cătălin Boja,
Ec. Silviu Niculescu, Dr. Gheorghe Noşca, Ing. Gabriela Zamfir, Tehn. Valentin Burcea, Drd. Robert Enyedi,
Asist. univ. Adrian Pocovnicu, Drd. Iulian Rădulescu,
Asist. univ. Cristian Toma, Stud. Adrian Vişoiu, Academia de Studii Economice Bucureşti, Prof. dr. Constantin Duguleană,
Prof. dr. Liliana Duguleană, Universitatea Transilvania Braşov
Prof. dr. Victor Ploae, Asist. univ. Ionuț Antohi,
Asist. univ. Silvia Ghiță-Mitrescu, Asist. univ. Irina Mădăraş,
Asist. univ. Cătălin Ploae, Universitatea Ovidius Constanța

Contractul de cercetare având ca temă *Sistem de evaluare a entităților bazate pe text* care se derulează în cadrul Programului Național de Cercetare-Dezvoltare şi Inovare INFOSOC are ca obiectiv realizarea unui sistem de evaluare a calității entităților text implementat cu un sistem de programe construit. Aceasta presupune parcurgerea următoarelor etape:

- definirea contextului, a entităților orientate pe text, a caracteristicilor de calitate;
- construirea indicatorilor asociați caracteristicilor de calitate a entităților orientate pe text, elaborare algoritmi şi proceduri;
- analiza calitativă şi validarea metricilor;
- definirea proceselor de evaluare şi implementare metrică;
- generalizare rezultate şi diseminare;
- auditarea rezultatelor.

În mod corespunzător, cercetarea întreprinsă îşi propune finalizarea prin rapoarte de cercetare a activităților derulate după cum urmează:

- Entități text şi caracteristici de calitate;
- Implementare metrici ale entităților text;
- Tehnici şi metode de validare metrici ale entităților text;
- Procese de evaluare a metricilor;
- Clase de metrici orientate text;
- Raport de auditare.

Corespunzător primei etape care se derulează în intervalul 1.10.2004 – 14.12.2004 şi care cuprinde activitățile:

- definirea inputurilor, definirea problemei de evaluare a entităților orientate pe text;
- realizarea unui studiu privind tipologiile de entități orientate pe text;
- analiza comparata a instrumentelor destinate evaluării de texte prin vocabular;
- identificarea caracteristicilor de calitate specifice fiecărui tip de entitate orientată pe text.

În mod corespunzător, cercetarea întreprinsă de către echipele din Academia de Studii Economice din Bucureşti, Universitatea Transilvania din Braşov şi Universitatea Ovidius din Constanța se concretizează prin elaborarea unor sinteze şi studii după cum urmează:

- sinteza documentara privind entitățile bazate pe text;
- clase de entități bazate pe text;
- studiu privind funcțiile si caracteristicile instrumentelor destinate implementării unor tehnologii de evaluare a calității entităților bazate pe text;
- studiu privind caracteristicile de calitate – completitudine, consistența, ortogonalitate, comparabilitate, eligibilitate, complexitate – entități orientate text.

Toate acestea sunt reunite în prezentul raport de cercetare, structurat pe zece capitole în care se definesc concepte de bază, se definesc caracteristicile de calitate, are loc definirea conceptului de ortogonalitate a datelor, se realizează prezentarea de evaluări folosind o bază de texte.

Ortogonalitatea datelor este o caracteristică importantă de calitate a acestora, asigurând identificarea în mod unic a unui text într-o mulțime considerată prin intermediul conceptului de amprentă.

Stabilirea amprentei unui text sau a unei aplicații informatice, în particular, determină identificarea în mod unic a acestuia în mulțimea textelor redactate, respectiv a programelor realizate şi lansate pe

piață de diverşi producători. Totodată, presupune un proces laborios de analiză a structurii acestuia, aducerea la formă interpretabilă din punct de vedere matematic a conținutului prin intermediul operației de normalizare.

Tratarea textelor şi programelor sursă ca date generează o nouă viziune în analiza aplicațiilor informatice, întrucât se aplică o aceeaşi măsură pentru elemente (software şi seturi de date) care numai aparent sunt diferite. Şi unele şi celelalte au ca suport structuri de date cu grade diferite de interdependență.

Există o corelație strânsă între caracteristica de ortogonalitate şi alte caracteristici de calitate ale software şi ale datelor. Studierea sistematică a corelațiilor dintre acestea va permite stabilirea momentului optim de înlocuire în aplicații informatice a componentelor indiferent de natura acestora.

În capitolul **Concepte de bază** sunt prezentate conceptele şi abordările teoretice din literatura de specialitate specifice entităților bazate pe texte.

Alfabetul A este o mulțime finită formată din n simboluri a_1, a_2, …. a_n.

Cuvântul este o succesiune formată din simboluri dispuse unul după celălalt. Un cuvânt c_j este caracterizat prin lungimea cuvântului, $L(c_j)$ exprimată ca număr de caractere ce intră în alcătuirea cuvântului.

Vocabularul V_A este o mulțime formată din cuvinte diferite. Lungimea vocabularului V_A, notată $Lgv(V_A)$ indică numărul de cuvinte care intră în alcătuirea vocabularului.

Separatorul este un simbol ce nu aparține alfabetului A, care are rolul de a delimita cuvintele ce alcătuiesc o secvență de cuvinte. Dacă se definesc separatorii s_1, s_2, …., s_r, oricare dintre ei se utilizează pentru a delimita cuvintele dintr-o înşiruire de cuvinte. Dacă un separator este utilizat consecutiv, se obține o secvență de separatori.

Textul T este o succesiune de cuvinte din vocabularul V_A separate prin simboluri speciale numite şi separatori. Lungimea textului Lgt(T) exprimă numărul de cuvinte din care este format textul. Lungimea textului Lgts(T) exprimă numărul de simboluri care intră în alcătuirea textului T.

Normalizarea textului T este operația prin care secvențele de separatori diferiți sunt înlocuite cu un separator considerat de bază.

Vocabularul textului este constituit din mulțimea cuvintelor diferite care apar într-un text.

Vocabularul textului VT este inclus în vocabularul V_A. Uneori, VT este identic cu V_A.

Frecvența de apariție a cuvântului c_j, notată f_j arată numărul de apariții ale cuvântului c_j în textul T.

Frecvența de apariție a simbolului a_i în textul T arată numărul de apariții al acestui simbol şi se notează cu g_i.

Entitățile bazate pe texte sunt construcții formate din şiruri de cuvinte caracterizate prin poziții ale cuvintelor în text, prin grupare de cuvinte pentru a defini un context, prin punerea în corespondență a cuvintelor cu elemente din lumea reală, cu acțiuni din lumea reală, cu fenomene din lumea reală şi cu atribute de ordin calitativ care grupează aspecte concrete din lumea reală în colectivități omogene în raport cu criterii stabilite.

Matricea de precedențe P are un număr de linii şi coloane egal cu numărul de cuvinte ce intră în alcătuirea vocabularului V_A. Elementul de pe linia i şi coloana j arată frecvența de apariție a perechilor de cuvinte (c_i, c_j).

Subvocabularul este o parte a vocabularului construit astfel încât intersecția oricărei perechi de subvocabulare are ca rezultat mulțimea vidă. Subvocabularele sunt mulțimi disjuncte.

Şablonul este o formalizare a unor reguli de construire a textelor. Textul este împărțit în subşiruri de cuvinte, iar şablonul impune ca fiecare subşir să aparțină unui subvocabular.

Textul structurat este rezultatul aplicării unui şablon sau unei reuniuni de şabloane.

Prototipul include cuvinte fixate în text şi cuvinte care se generează şi se intercalează între cuvintele fixate, conducând la o construcție proprie a unui text.

Identificarea şi construirea sistemului de clasificare a textelor reprezintă un pas important şi baza definirii de **Structuri de entități**, care constituie un alt capitol al acestui raport de cercetare. Sunt trecute în revistă contribuțiile din literatura de specialitate, definirea şi construirea de structuri de entități bazate pe texte.

Noțiunea de text structurat este în strânsă legătură cu proiectele structurate prin prisma conținutului proiectelor. Proiectele conțin texte oarecare, figuri, figuri şi text, tabele cu date. În schimb, proiectele structurate conțin elementele impuse de protocolul de desfăşurare al concursului. Anumite elemente pot sa nu se regăsească sau pot fi conține în totalitate, dar toate proiectele structurate au un element comun: textul. Toate proiectele structurate conțin cel puțin doar text, de unde şi noțiunea de text structurat.

Textul T este text structurat dacă este format din subtexte. Practic textul T este privit ca fiind alcătuit dintr-o concatenare de texte. Proprietatea definitorie a textelor structurate este aceea că fiecare subtext al textului T este privit ca o entitate de sine stătătoare. Fiecare subtext are propriul vocabular, lungime proprie, existând posibilitatea analizei comparate a respectivului subtext cu un alt subtext din alt text sau din textul T.

Comunicarea interumană este coerentă, adică sunt comunicate o serie de idei într-o ordine logică şi într-o formă eligibilă pentru ambii parteneri ai comunicării. Lipsa coerenței duce la bariere în comunicare, ceea ce produce efecte neaşteptate.

În comunicarea în scris coerența este asigurată prin interdependența textelor. Un text este considerat coerent dacă subtextele din care este compus au elemente comune, adică textul tratează ceea ce şi-a propus. Dacă nu este asigurată această proprietate textul respectiv este compus din texte independente(nu au elemente în comun) ceea ce ne duce la concluzia că respectivul sau respectivele texte au fost introduse doar pentru a îndeplinii condiția numărului de pagini. În schimb dacă toate subtextele care compun textul structurat T sunt interdependente atunci textul T este considerat coerent. Totuşi, există anumite situații în care ofertantul impune utilizarea textelor independente.

Coerența textelor este clasificată după cum urmează:

- coerență nulă;
- coerență simplă;
- coerență compusă;
- coerență multiplă.

Caracteristici de calitate reprezintă următorul capitol al raportului de cercetare în care sunt prezentate principalele caracteristici şi elementele lor specifice, cu ajutorul cărora se diferențiază entitățile bazate pe texte, în vederea ierarhizării lor pentru a obține calificative, pentru atribuire de fonduri, pentru inserarea în clase omogene.

Entitățile text sunt construcții foarte generale, iar entitățile-proiect sunt exemplificări pentru acestea, fiind uşor de manipulat întrucât condițiile de calitate impuse entităților text se regăsesc în totalitate în mulțimea entităților-proiect, în general, şi, respectiv, în mulțimea entităților-proiect TIC.

Caracteristica de ***eligibilitate*** a entității-proiect TIC este de importanță majoră, întrucât include cerințe strict legate de:

- calitatea echipei de executanți a proiectului, din punct de vedere a obiectului de activitate;
- poziția în plan financiar şi juridic, nefiind sancționată sau neavând incapacitate de atragere resurse;
- experiența în domeniu şi rezultatele obținute în derularea proiectelor anterioare;
- capacitatea de atrage resurse, de a utiliza resurse şi de a le gestiona;
- existența capacității manageriale a proiectului în ansamblu.

Structurabilitatea proiectului este caracteristica de calitate prin care se creează evaluatorilor posibilitatea de a urmări distinct părți ce definesc cerințele impuse prin program.

Claritatea proiectului este dată de folosirea în capitolul CK_i de cerințe cheie incluse în lista de priorități a programului de finanțare şi de introducerea de cuvinte cheie proprii. Din aproape, în aproape capitolul CK_i utilizează cuvinte cheie ale capitolelor precedente şi nu face referire la cuvinte cheie care sunt definite ulterior.

Claritatea presupune o listă de acronime care definesc acronimele utilizate, un glosar de termeni pentru definirea de procese, de materiale, produse, fenomene, unități de măsură cu rol de a elimina ambiguități asupra expresiilor cantitative întrebuințate în text.

Consistența ofertei de entități-proiect TIC este dată de prezența în cadrul structurii din capitolul CK_i a elementelor de descriere D_{i1}, D_{i2}, …, D_{imi}, ce reprezintă activități, materiale, echipamente, persoane, evenimente, servicii aflate în ordine. Între elementele de descriere există relația de disjuncție:

$$D_{ij} \cap D_{ik} = \phi$$

ceea ce presupune că acestea nu au caracter repetitiv de preluare de la un element la altul.

De asemenea elementul D_{ik} nu trebuie să anuleze prezența $Di_{j,}$ k> j, sau D_{ik} nu trebuie să fie negația lui D_{ij}, adică $\overline{D_{ij}}$.

Consistența unui proiect este analizată atât în cadrul fiecărui capitol, cât şi între capitole. Ofertantul de entități-proiect TIC enumeră cerințe ale tehnologiilor utilizate, fiind obligat să achiziționeze

instrumente capabile să implementeze aceste tehnologii şi să angajeze personal cu calificare de a utiliza instrumentele.

Completitudinea este dată de abordarea sub forma structurii arborescente a textului. În cazul structurii arborescente se definesc funcțiile *in()* şi *aut()* care evidențiază arcele incidente spre interior, respectiv, arcele incidente spre exteriorul unui nod.

Realismul abordării reprezintă un element de bază în lista caracteristicilor de calitate a ofertei, element important mai ales pentru managementul calității proiectului. Realismul este dat de modul cum este ales titlul, de obiectivul formulat şi de lungimile listelor cu care se construiesc tabelele.

Realismul managerului de proiect este determinat de modul în care formează echipa, de cum derulează cererile de tranşe de finanțare, de cum eşalonează activitățile.

Ortogonalitatea entităților-proiect, *H( )*, este utilizată pentru analiză de text între capitole. Două capitole CK_i şi CK_j sunt ortogonale dacă textele lor nu au elemente identice şi ortogonalitatea are valoarea:

$$H(CK_i, CK_j) = 1$$

Sunt situații în care ofertanții de proiecte preiau în partea de concluzii CK_n textul introducerii CK_1. În acest caz ortogonalitatea are valoarea:

$$H(CK_1, CK_n) = 0$$

În managementul programului de finanțare TIC este important ca structurile de entități-proiect PP_1, PP_2, …, PP_{mp} să fie ortogonale, adică

$$H(PP_i, PP_j) = 1$$

făcându-se referire la colecția de oferte a proiectelor pe toată durata programului, ca reuniune de proiecte lansate în cadrul licitațiilor.

O astfel de abordare elimină obținerea de finanțări destinate pentru aceeaşi ofertă. Sunt situații în care programele de finanțare vizează ca prin ofertă să se obțină prototipuri adaptabile. Un proiect care a obținut finanțarea pentru o astfel de temă, evident, gestionează variantele generate de prototip, finanțările fiind adecvate, fără a trata oricare dintre variante ca un nou produs, diferit de cele existente.

Se realizează dezvoltarea unei tehnici de analiză pe structuri de text cu aplicații la proiecte.

Gradualitatea ofertei de entități-proiect TIC vizează creşterea nivelului de detaliere de la un capitol la altul. La început este definit obiectivul, iar după aceea sunt prezentate tehnologiile existente, efectuându-se o analiză comparată.

Descriere graduală a ofertei de entități-proiect TIC implică parcurgerea etapelor:

- se alege tehnologia de dezvoltat;
- se prezintă în detaliu activitățile tehnologiei alese;
- se descriu resursele necesare, fie că sunt de reprezentate de echipamente, personal sau resurse financiare;
- se construiesc tabele cu consumuri, costuri, resurse care să explice nivelurile agregate;
- se evaluează costul total al proiectului.

Corectitudinea proiectelor constă în includerea de texte care sunt acceptate ca fiind în concordanță cu elementele de bază din domeniul pentru care se efectuează finanțarea. Corectitudinea vizează denumiri de proces, tehnologie, operații, utilizarea de concepte, prezentarea de modele, semnificația variabilelor. Se respectă rezultatele validate de practica existentă în literatura de specialitate. De asemenea este vizată respectarea deontologicii privind ceea ce există şi ceea ce se adaugă, fiind privit ca elemente de noutate, respectând cerințele legii dreptului de autor.

Corectitudinea vizează citările, referințele bibliografice, structura modelelor definițiilor şi modul în care sunt utilizate extrase din lucrări publicate.

Caracteristica este strâns legată de modul logic de eşalonare a activităților, de nivelul consumurilor de resurse, de estimările efectuate. Este important modul în care se realizează descrierea tehnologiilor sau a operațiilor, completare tabelelor (A,R), (A,T) şi (R, T) prin includerea de niveluri reale sau niveluri medii.

Dacă este dată în mod transparent procedura de evaluare a ofertelor de proiecte, corectitudinea vizează în special autoevaluarea cu obținerea unei diferențe minime față de rezultatul evaluării.

Definirea caracteristicii de calitate privind **Ortogonalitatea entităţilor** se realizează prin intermediul diferenţelor de vocabulare, de frecvenţe în utilizarea cuvintelor, în diferenţe de lungimi de texte.

Conceptul de ortogonalitate a elementelor este deosebit de important în contextul constituirii listei de entităţi cât mai diferite între ele.

Două drepte sunt ortogonale dacă unghiul format la intersecţia acestora are cosinusul egal cu zero, cu alte cuvinte cele două drepte sunt perpendiculare. Un ansamblu de drepte este ortogonal dacă dreptele ce-l compun sunt perpendiculare două câte două .

Două plane sunt ortogonale dacă unghiul format la intersecţia lor are cosinusul egal cu valoarea zero, adică cele două plane sunt perpendiculare. Un ansamblu format din mai multe plane este ortogonal dacă planele ce-l formează sunt ortogonale două câte două.

Doi vectori sunt ortogonali dacă produsul scalar al acestora este nul.

În continuare o dată este reprezentată uzual ca un număr (125 sau 0 sau –400.72 sau +25.3e-4) sau un şir de caractere (“maşina” sau “125” sau “Cluj-Napoca-2000”).

Extinzând, rezultă că datele D_1 si D_2 sunt ortogonale **semantic**, dacă conţinutul informaţional al acestora, sensul lor, diferă intr-o manieră categorică şi **semiotic**, dacă acestea au o formalizare matematică total diferită.

Se consideră vocabularul $V_A = \{c_1, c_2, \ldots, c_n\}$, unde $c_1, c_2, \ldots, c_n$ sunt cuvinte din vocabular.

Se construiesc frazele:

F_1 formată din cuvintele $c_{k1}, c_{k2}, \ldots, c_{kn}$

F_2 formată din cuvintele $c_{l1}, c_{l2}, \ldots, c_{lm}$

Lungimea ca număr de cuvinte a acestor fraze este:

$Lg(F_1) = Lg(c_{k1}\ c_{k2} \ldots c_{kn}) = p$ cuvinte

$Lg(F_2) = Lg(c_{l1}\ c_{l2} \ldots c_{lm}) = m$ cuvinte

Frazele F_1 si F_2 sunt ortogonale dacă ele diferă atât ca lungime cât şi din punct de vedere al conţinutului. Nu există nici un cuvânt din fraza F_1 care să se regăsească în fraza F_2 şi invers.

Pentru colectivitatea A*A*, elementele aa_1, aa_2 , … , aa_n , diferite, se descriu printr-un număr de *m* caracteristici specifice colectivităţii, C_1 , C_2 , … , C_m.

Pentru două elemente oarecare aa_i şi aa_j aparţinând colectivităţii *AA* se înregistrează nivele ale caracteristicilor după cum urmează:

$aa_i = (C_{i1}, C_{i2}, \ldots, C_{im})$

$aa_j = (C_{j1}, C_{j2}, \ldots, C_{jm})$

unde C_{ik} este nivelul caracteristicii C_k pentru elementul aa_i .

Un element *x* este ortogonal pe colectivitatea *AA* dacă şi numai dacă $<x, aa_i> = 0$, oricare ar fi aa_i, element al colectivităţii.

În spaţiul R^n, pentru elementele x şi y aparţinând acestuia se defineşte produsul scalar dintre cele două elemente prin $<x,y> = x_1*y_1+x_2*y_2+\ldots+x_n*y_n$, dacă $x= (x_1, x_2 \ \ldots \ ,x_n)$ şi $y = (y_1 ,y_2 \ \ldots \ , y_n)$.

Se defineşte operaţia Θ aşa fel încât C_{ik} Θ C_{jk} are valoarea zero dacă pentru caracteristica C_k nu există valori comune elementelor aa_i şi aa_j şi valoarea 1 în caz contrar.

Se defineşte pseudoprodusul scalar prin $<<aa_i , aa_j >>$.

Elementul aa_i este ortogonal cu elementul aa_j dacă $<<aa_i, aa_j>> = 0$, dacă (C_{i1} Θ C_{j1} + C_{i2} Θ C_{j2} + … + C_{im} Θ C_{jm}) = 0.

În condiţiile în care elementele grupului sunt descrise prin caracteristici se consideră această relaţie adevărată dacă fiecare termen este egal cu zero: C_{i1} Θ C_{j1} = 0 , C_{i2} Θ C_{j2} = 0 , … , C_{im} Θ C_{jm} = 0.

Relaţia C_{ik} Θ C_{jk} = 0 , k aparţinând mulţimii {1, 2, …, m}, exprimă faptul că elementele aa_i şi aa_j ale grupului sunt total diferite după caracteristica C_k .

Se defineşte indicatorul care evidenţiază ortogonalitatea dintre două elemente ale unei colectivităţi sau, mai flexibil, care evidenţiază diferenţele existente între elemente.

Indicatorul normat I, cuprins în intervalul [0,1] are valorile:

I = 0, dacă elementele sunt ortogonale (nu au nimic în comun).

I = 1, elementele sunt identice (nu diferă prin nici o caracteristică).

Dacă valorile indicatorului I tind către 1 înseamnă că seturile de date tind către ortogonalitate, iar dacă valorile indicatorului normat I sunt apropiate de 0 înseamnă că seturile de date au foarte multe elemente identice.

Indicatorul se defineşte pentru măsurarea gradului de ortogonalitate dintre două seturi de date, entităţi bazate pe texte şi, de asemenea, pentru ortogonalitatea unui număr oarecare de seturi de date, respectiv entităţi bazate pe texte.

Capitolul **Amprente** are menirea de a introduce un nou concept cu largă utilizare în studiul calitativ al entităților bazate pe texte, care se fundamentează pe eşantionare şi pe extensie de rezultate de la eşantion către colectivitatea de bază.

Pentru construirea eşantionului se pot folosi mai multe procedee, şi anume:

- procedee de eşantionare aleatoare: loterie, numere aleatoare, mecanic;
- eşantionare dirijată;
- eşantionare mixtă.

Pentru determinarea volumului eşantionului se efectuează următoarele operații:

- se estimează dispersia mulțimii din care se face eşantionarea;
- se stabileşte probabilitatea cu care eşantionul va fi reprezentativ pentru mulțime inițială;
- i se asociază probabilității stabilite valoarea tabelară a variabilei *Student*;
- se stabileşte eroarea maxim admisă.

Mecanismele de obținere a şirurilor de numere pseudoaleatoare aparținând unui interval constă în împărțirea intervalului inițial într-un număr de subintervale egal cu numărul de valori ce trebuie generate. Se realizează punerea în corespondență a subintervalului căruia aparține numărul pseudoaleator generat cu valoarea din mulțimea de bază formată din kinv elemente b_1, b_2, ..., $b_{kinv.}$

Algoritmul care construieşte amprenta textului utilizând numere pseudoaleatoare realizează următoarele:

- o citeşte textul;
- o stabileşte lungimea textului;
- o numără cuvintele separate prin spațiu, punct, virgulă;
- o stabileşte lungimea vocabularului;
- o stabileşte lungimea alfabetului;
- o generează vectorul cu numere pseudoaleatoare;
- o extrage eşantionul de caractere;
- o calculează frecvențele de apariție ale elementelor din alfabet;
- o extrage eşantionul de cuvinte;
- o construieşte vocabularul eşantionului;
- o calculează frecvențele de apariție a elementelor de vocabular;
- o calculează numărul inversiunilor necesare sortării frecvențelor**;**
- o normează numărul inversiunilor **;**
- o generează indicatorii xa_1, xa_2, xa_3, $xa_{4;}$
- o calculează indicatorul agregat GA;
- o afişează valorile următorilor indicatori:
  - coeficientul de variație al frecvențelor caracterelor din alfabetul eşantionului;
  - coeficientul de variație al frecvențelor cuvintelor din vocabularul eşantionului;
  - numărul inversiunilor necesare sortării frecvențelor caracterelor din alfabetul eşantionului, normat;
  - numărul inversiunilor necesare sortării frecvențelor cuvintelor din vocabularul eşantionului, normat;
  - indicatorul agregat GA.

Capitolul **Operații pe entități text** stabileşte modalități de efectuare a operațiilor pe entități text care conduc la construirea de entități text echivalente, la compuneri de entități şi la obținerea de noi entități prin procese de agregare.

I Copierea de fişiere

II Conversia de fişiere

III Translatarea fişierelor sursă

IV Transformări pe text sursă

V Restructurarea programelor

VI Schimbarea structurilor de date utilizate

VII Omogenizarea tipurilor de funcții

Agregarea entităților text:

Concatenare de entități text reprezintă o operație des întâlnită în cazul în care echipe de specialişti elaborează independent capitole ale unor lucrări.

Compunerea prin inserare apare în cazul entităților când se realizează citări pentru asigurarea rigurozității preluărilor din surse bibliografice. În cazul produselor software macroapelurile reprezintă inserări şi apelurile de proceduri reprezintă referiri.

Agregarea prin extragere subşiruri cu aceeaşi poziție corespunde situației în care se asociază definiții echivalente sau se introduc în programe texte ASM pentru a gestiona mai eficient resursele.

Extragere şiruri de subşiruri cu poziții diferite corespunde situației în care în procesul de mentenanță software anumite module trebuie dezactivate, iar în cazul unor monografii pentru a gestiona redundanța se elimină capitole sau paragrafe descrise într-o manieră mai substanțială anterior.

Evaluarea entităților text reprezintă partea de cercetare care are o tratare specială în cadrul etapei a treia din proiectul *Sistem de evaluare a entităților bazate pe text,* acum fiind definit contextul evaluării în ideea de fundamentare a sistemului caracteristicilor de calitate ce va fi utilizat în procesele de evaluare.

Managementul calității entitate text presupune decizii succesive în timpul elaborării ofertelor. Există un proces de autoevaluare care se bazează pe analiza textului elaborat şi pe măsurarea unor indicatori. Procesul precede evaluarea ofertelor de către specialişti în vederea ierarhizării pentru a aloca fondurile de finanțare. Este vorba de evaluarea ofertantului unei propuneri în vederea efectuării de corecții înaintea depunerii acesteia la unitatea de management a programului de finanțare.

Pentru celelalte caracteristici de calitate a entității *Prt* se stabilesc o serie de ipoteze în raport cu care sunt definiți indicatori de măsurare a nivelurilor respectivelor caracteristici.

În primul rând, indicatorii se construiesc în aşa fel încât valorile lor să aparțină intervalului [0; 1]. Dacă un indicator are valoarea calculată egală cu zero, înseamnă că entitatea text pentru care s-a calculat indicatorul, este lipsită de caracteristica de calitate pentru care este asociat indicatorul.

În cazul în care nivelul calculat al indicatorului este egal cu 1 sau apropiat de această valoare, înseamnă că entitatea posedă însuşirile cerute de caracteristica de calitate căruia indicatorul i-a fost asociat.

În al doilea rând, pentru colectarea de date sunt luate în considerare elemente care să asigure reproductibilitatea procesului de măsurare a caracteristicii de calitate indiferent de condiții şi de echipa care efectuează măsurătoarea.

În al treilea rând, metricile definite trebuie să aibă caracter operațional în sensul alocării de resurse cu un nivel acceptabil pentru culegerea de date, pentru efectuarea calculelor şi pentru interpretarea rezultatelor. Costurile aferente trebuie să fie suportabile.

În al patrulea rând, se impune o transparență totală în derularea de procese de validare a metricelor înaintea efectuării evaluării şi înaintea stabilirii pragurilor de acceptabilitate a entităților.

Luarea deciziilor şi acceptarea acestora de către ofertanți elimină procesele de reevaluare a întregii entități atunci când se schimbă criteriile de ierarhizare pe parcursul procesului de evaluare.

În al cincilea rând, se stabilesc criteriile şi caracteristicile de evaluare, precizându-se ponderile. În procesul de transparență, pentru fiecare dintre caracteristici se atribuie puncte şi totalul punctelor este cuprins între zero şi o sută de puncte. Ofertanții trebuie să aibă informații asupra punctelor acordate caracteristicilor de calitate şi asupra formulelor de calcul.

Pentru operaționalizarea conceptului de entitate text, în capitolul **Baze de entități text** se procedează la stabilirea unei liste de entități text care îndeplinesc condiții de ortogonalitate, considerată

listă de referință, iar procesele de adăugare, de inserare, de concatenare se produc dacă şi numai dacă noua listă rezultată îşi menține caracterul de listă formată din elemente de referință.

Economia modernă utilizează numeroase tehnici şi metode de analiză, dintre care cele orientate spre latura cantitativă a evoluției proceselor, sunt cele mai importante.

Construirea de entități text are ca obiectiv prezentarea aspectelor esențiale şi a factorilor de influență, elaborarea de modele, construirea de ipoteze, enumerarea listelor de activități şi de resurse, realizarea de aprecieri calitative privind dinamica fenomenelor.

Există o mare diversitate de entități, fiecare din ramurile ştiințelor rezervând spații de prezentare deosebit de largi rezultatelor obținute pe baza cercetărilor şi prezentărilor prin articole, rapoarte, studii, monografii, tratate, comunicări, oferte.

Se identifică următoarele tipuri de entități: entități directe, entități bazate pe modele de optimizare, entități ce conțin elemente de neliniaritate.

Practica economico-socială presupune manipularea cu numeroase tipologii de entități text, a căror prezentare este realizată în capitolul **Diversitatea entităților**. Aceasta presupune stabilirea unor măsuri pentru similaritate la nivel de cuvânt, la nivel de şir de cuvinte, la nivelul întregii entități.

Sistemul de indicatori pentru măsurarea diversității trebuie să îndeplinească cerințele de senzitivitate, necatastroficitate, necompensator.

Entitățile text se prezintă sub o foarte mare diversitate de construcții. De aceea, exemplificările acestui raport de cercetare conțin texte din limbaje artificiale date spre exemplificare, texte din literatura de specialitate, precum şi texte sursă ale unor programe scrise în limbajul C/C++.

Abordarea din raportul de cercetare permite extensii spre texte în care simbolurile alfabetului sunt hieroglife, alte reprezentări grafice, reprezentări prin culori, precum şi diagrame utilizate în caietele de regie de teatru, operă sau de balet.

Raportul de cercetare se încheie cu **Concluzii** în care sunt sistematizate rezultatele cercetării din etapa întâia a proiectului *Sistem de evaluare a entităților bazate pe text*, având conținut definirea contextului, a entităților orientate pe text şi a caracteristicilor lor de calitate.

Conținutul acestui raport este rezultatul unei activități de cercetare ştiințifică interuniversitară, întreprinsă de cadre didactice de la Catedra de Informatică Economică a Academiei de Studii Economice din Bucureşti, de la Universitatea Transilvania din Braşov şi de la Universitatea Ovidius din Constanța, fiind concretizată şi prin prezentarea şi publicarea unor soluții obținute în intervalul aferent acestei etape la manifestări ştiințifice din țară şi străinătate, precum şi la reviste de specialitate din țara noastră.

Bibliografia întocmită pentru această sinteză conține lucrări de specialitate din literatura română şi străină. Unele din titluri sunt elaborate de membrii echipei de cercetare.

[AMIT03] IVAN, I., AMITROAIE, M., *Building text's fingerprint*, Economy Informatics, vol. 3, nr. 1, 2003, pg. 57 – 62

[APOS03] IVAN, I., APOSTOL, C., *Certificarea produselor program prin amprente*, Revista Română de Automatică şi Informatică, vol. 13, nr. 1, 2003, pg. 32 – 38

[BALG03] BALOG, A., şi colectiv – *Specificarea sistemului de metode şi proceduri de măsurare şi evaluare a calității produselor / serviciilor*, Raport de cercetare, Programul CALIST, Faza 2, iunie 2003

[BALO04] BALOG, A., (Ed.) – *Calitatea sistemelor interactive*, Bucureşti, Editura Matrix Rom, 2004

[BATE00] BATES, J., TOMPKINS ,T., - *Utilizarea Visual C++ 6* , Bucureşti, Editura Teora, 2000

[BIBA82] BARON, T. IVAN, I., BALOG, A., *Evaluarea calității produselor program,* Revista Economică, supliment nr. 17, 1982, pg. 4 – 6

[BIBU00] BIBU, N., BRANDAS, C., – *Managementul prin proiecte*, Timişoara, Editura MIRTON, 2000

[BLGO00] BALOG, A., *Calitatea produselor software. Măsurare, analiză şi evaluare,* Bucureşti, Editura INFOREC, 2000

[BLOG97] B ALOG, A., – *Analiză statistică şi evaluarea calității software-ului,* Bucureşti, Editura Calypso, 1997

[BODE00] BODEA, C., N., BODEA, V., ÎNTORSUREANU, I., POCATILU, P., LUPU, R. A., COMAN, D., *Managementul proiectelor*, Bucureşti, Editura INFOREC, 2000

[BOGU98] BOGURAEV, B., KENNEDY, C., *Salience-Based Content Characterisation of Text Documents*, IBM Thomas J Watson Research Center, USA, 1998

[BOJA02] IVAN, I., NICULESCU, S., BOJA, C., *Clonarea bazelor de date*, Revista Română de Automatică şi Informatică, vol. 12, nr. 4, 2002, pg. 46 – 53

[CAPI03] IVAN, I., POPA, M., CAPISIZU, S., FLORESCU, B., IVAN, L., – *Characteristics of the Informatics Clones*, în „Master of International Business Informatics Handbook", Bucureşti, Editura ASE, 2003, pg. 207 – 223

[DODM94] Departmet of Defence 8320.1-M – *Data Quality Assurance Procedures (Draft). Quality Information for a Strong Defence,* 1994

[HALS77] HALSTEAD, M. H., *Elements of Software Science*, Elsevier – North Holland, Amsterdam, 1977

[IARH84] IVAN, I., ARHIRE, R., MACESANU, M., *Program Complexity Analysis, Hierarchy, Classification*, SIGPLAN NOTICES, vol. 22, nr. 4, 1984, pg. 94 – 102

[INOP96] IVAN, I., NOŞCA, G., PÂRLOG, A., *Asigurarea calității datelor,* în „Asigurarea calității", an 2, nr. 8, 1996, pg. 8 – 15

[INPO03] IVAN, I., POPA, M., POCATILU, P., *The Fingerprint – an Unique Way to Identify Programs*, Proceedings of the International Symposium "Knowledge Technologies in Business and Management", Iaşi, 6 iunie 2003, pg. 40 – 45

[IOIP04] IVAN, I., POPA, M., *Ortogonalitatea produselor program orientate obiect*, în „Informatica Economică", vol. 8, nr. 4, 2004, în curs de apariție

[IOPO03] IVAN, I., POCATILU, P., POPA, M., *Ortogonalitatea – caracteristică de calitate a proiectelor regionale*, Comunicare în al III-lea simpozion național al ARSR "Avantaje competitive şi dezvoltare regională", Bucureşti, 22 – 23 mai 2003

[IPFC03] IVAN, I., POPA, M., FLORESCU, B., CAPISIZU, S., *Aspecte juridice ale clonării informatice*, în revista "Informatica Economică", vol. 7, nr. 1, Bucureşti, 2003, pg. 46 – 50

[IPOC02] IVAN, I., POCATILU, P., POPA, M., SACALĂ, M., *Clonarea informațională*, în "Revista Română de Informatică şi Automatică", vol. 12, nr. 2, Bucureşti, 2002, pg. 47 -52

[IPOP04] IVAN, I., POPA, M., POPESCU, A., *The Aggregation of the Text Entities*, în „Studii şi Cercetări de Calcul Economic şi Cibernetică Economică", 2004, în curs de apariție

[IPOS02] IVAN, I., POCATILU, P., POPA, M., SACALĂ, M., *Information Cloning*, Proceedings of the International Symposium Regional Problems in the Context of Globalization Process, Chişinău, Moldova Republic, 9 – 10 octombrie 2002, pg. 371 – 375

[ITOV82] IVAN, I., TOVISSI, L., MOSCOVICI, E., *Utilizarea analizei entropice in studiul complexității programelor cu aplicații la normarea activității de programare*, Buletinul Român de Informatică, nr. 4, 1982, pg. 39 – 44

[IVAN02] IVAN, I., POCOVNICU, A., *QSPM Software pentru managementul calității proiectelor*, revista Informatică Economică, vol. 6, nr. 4, 2002, pg. 67-70.

[IVAP04] IVAN, I., POPA, M., *The Analyse Based on Texts Fingerprint*, Proceedings of „The Central and East European Conference in Business Information Systems", Cluj-Napoca, 20 – 22 mai 2004, pg. 554 – 561

[IVAT99] IVAN, I., TCACIUC, S., *Spre managementul total al calității datelor,* Revista Română de Statistică, 1999

[IVAV04] IVAN, I., VIŞOIU, A., POPA, M., *Ortogonalitatea – caracteristică a calității bazei de modele economice*, în „Revista Română de Informatică şi Automatică", vol. 14, nr. 3, 2004, pg. 89 – 100

[IVBO03g] IVAN, I., POCATILU, P., POPA, M., BOJA, C., NICULESCU, S., *Replicarea bazelor de date*, lucrare în simpozionul "Tehnologii educaționale pe platforme electronice în învățământul ingineresc" al Universității Tehnice de Construcții, Bucureşti, Editura CONSPRESS, 9 – 10 mai 2003 (în format electronic)

[IVBO04] IVAN, I., BOJA, C., *Metode statistice în analiza software*, Bucureşti, Editura ASE, 2004

[IVCA03] IVAN, I., POPA, M., CAPISIZU, S., BREDA, L., FLORESCU, B., *Clonarea informatică*, Bucureşti, Editura ASE, 2003

[IVNP99] IVAN, I., NOŞCA, G., PÂRLOG, O., CĂCIULĂ, R.,

TCACIUC, S., *Calitatea datelor*, Bucureşti, Editura INFOREC, 1999

[IVPA03] IVAN, I., POPA, M., AVRAM, C., *Utilizarea gradului de similaritate în construirea amprentei software*, în revista ”Studii şi Cercetări de Calcul Economic şi Cibernetică Economică”, vol. 37, nr. 2, Bucureşti, 2003, pg. 39 – 54

[IVPD04] IVAN, I., POPA, M., DRĂGOI, R., *Validarea ortogonalității firmelor şi emblemelor agenților economici*, în „Studii şi Cercetări de Calcul Economic şi Cibernetică Economică”, vol. 38, nr. 3, 2004, pg. 17 – 24

[IVPF03] IVAN, I., POPA, M., FLORESCU, B., UNGUREANU, D., *Juridical Aspects of the Program Generators Cloning*, Proceedings of the 34-th International Scientific Symposium of the Military Equipment and Technologies Research Agency, Volume I, Bucureşti, 29 – 30 mai 2003, pg. 283 – 288

[IVPO04] IVAN, I., POPA, M., *Tipuri de metrici ale textelor*, în „Studii şi Cercetări de Calcul Economic şi Cibernetică Economică”, vol. 38, nr. 1, 2004, pg. 25 – 36

[IVPP02] IVAN, I., POCATILU, P., POPA, M., *Analiza Cantitativă în managementul proiectelor*, în revista “Studii şi Cercetări de Calcul Economic şi Cibernetică Economică”, vol. 36, nr. 2, Bucureşti, 2002, pg. 9 – 21

[IVPS02] IVAN, I., POPA, M., SACALĂ, M., *Ortogonalitatea datelor*, în “Revista Română de Statistică”, anul 11, nr. 4, Bucureşti, 2002, pg. 30 - 45

[IVTO04] IVAN, I., POPA, M., TOMA, C., BOJA, C., *Data Metrics Properties*, Proceedings of the International Symposium „Innovative Applications of Information Technologies in Business and Management”, Iaşi, 22 – 23 octombrie 2004, pg. 45 – 56

[IVVI04] IVAN, I., VIŞOIU, A., *Generator de modele cu argument întârziat*, în „Revista de Comerț”, vol. 5, nr. 1, 2004, pg. 47 – 50

[KERZ96] KERZNER, H., *Project Management, A Systems Approach to Planning, Scheduling and Controlling*, John Wiley & Sons Inc., New York, 1996

[KETZ96] KRETZ, L., *Project Partners*, International Institute for Learning, Inc., 1996

[KRET96] KRETZ, L., *If you manage project managers*, International Institute for Learning, Inc., 1996

[KRTZ96] KRETZ, L., *Project Management & Leadership Concepts*,

International Institute for Learning, Inc., 1996

[MACE85] MĂCEŞANU, M., ARHIRE, R., IVAN, I., *Clase de complexitate pentru produse program*, în „Buletinul Român de Informatică”, nr. 1, 1985, pg. 63 – 68

[NICU04] NICULESCU, S., *Analiza ortogonalității proiectelor IT structurate*, Lucrare de licență, Bucureşti, ASE, 2004

[NOSC04] NOŞCA, G., IVAN, I., POPA, M., *A Text Fingerprints-Based Analysis Algorithm*, Proceedings of „The 5th Biennial International Symposium SIMPEC 2004”, Braşov, 14 – 15 mai 2004, pg. 281 – 286

[PARK01] PARK, Y., BYRD, R., BOGURAEV, B., *Automatic Glossary Extraction: Beyond Terminology Identification*, IBM Thomas J Watson Research Center, USA, 2001

[PARL04] PÂRLOG, O., *Modele de optimizare a calității datelor*, Bucureşti, ASE, 2004, teză de doctorat

[PCIV03] IVAN, I., POCOVNICU, A., *Text's Fingerprint Building, Using Sampling Base don Generation of Pseudo-Random Numbers,* Proceedings of International Workshop IE & SI, Timişoara, 23 – 24 mai 2003, pg. 25 – 28

[PITA96] PITAGORSKY, G., *Project Management, Process Handbook*, International Institute for Learning, Inc., 1996

[POCA03] IVAN, I., POCATILU, P., POPA, M., MIHAI, T., IVAN, L., *Data Orthogonality*, în „Master of International Business Informatics Handbook”, Bucureşti, Editura ASE, 2003, pg. 235 – 249

[POCO03] IVAN, I., POPA, M., POCOVNICU, A., *Software pentru evaluarea amprentei textului utilizând numere pseudoaleatoare*, în revista „Informatica Economică”, vol. VII, nr. 3, 2003. pg. 77 – 80

[POPI04] IVAN, I., POPA, M., TOMA, C., RĂDULESCU, I., *The Aggregation of the Data Orthogonality Metrics*, Proceedings of „The 35th International Scientific Symposium of METRA”, vol. 1, Bucureşti, 27 – 28 mai 2004, pg. 590 – 595

[PORO93] POROJAN, D., *Statistica şi teoria sondajului*, Bucureşti, Casa de editura şi presă Şansa SRL, 1993

[REDM92] REDMAN, T., *Data Quality: Management and Technology,* Bantam Books, 1992

[REME01] REMENYI, D., *Investițiile în TI, elaborarea unui studiu de eficiență*, Editura Club Europa, 2001

[SMEI98] SMEUREANU, I., IVAN, I., DÂRDALĂ, M., *Structuri şi obiecte în C++*, Bucureşti, Editura CISON, 1998.

[SMEU01] SMEUREANU, I., DÂRDALĂ, M., *Programarea în limbajul C/C++,* Bucureşti, Editura CISON, 2001

[STAC01] STANCIU, E., *Căi de creştere a gradului de reutilizare a componentelor software*, Bucureşti, ASE, 2001, referat de doctorat

[STAN00] STANCIU, E., *Reutilizarea componentelor program în sisteme complexe*, ASE Bucureşti, 2000, referat de doctorat

[TOMI04] IVAN, I., TOMA, C., POPA, M., VIŞOIU, A., *Analiza ortogonalității componentelor din baza de modele economice*, comunicare în „A IX-a Sesiune de Comunicări Ştiințifice a Cadrelor Didactice", Universitatea Româno-Americană, Bucureşti, 28 – 29 mai 2004

[****70] *I.B.M. - Application Program, Fifth Edition*, August 1970

[www] www.research.ibm.com