Entităţi bazate

Prof. univ. dr. *Ion IVAN*
Daniel MILODIN
Marius POPA
Cosmin LUGOJI

Rezumat

Articolul prezintă conceptul de entitate bazată şi se evidenţiază operaţii pe acest tip de entitate. Se definesc indicatori de apartenenţă şi se identifică metode şi căi de căutare. Se exemplifică evaluarea entităţilor bazate folosind software EBASOF.

Abstract

It is presented the based entity concept and operations on this kind of entity. There are defined appurtenance indicators and there are identified methods and ways for searching. The based entity evaluation is exemplified using the EBASOF software.

În managementul organizaţiilor se definesc strategii de management şi planuri de acţiune. Scopul este de a acţiona şi de a obţine rezultatele conform obiectivelor stabilite.

Pentru situaţii şi contexte similare este eficientă utilizarea de şabloane prin a căror instanţiere se obţin strategii şi planuri care sunt adaptate la situaţia de fapt. Instanţele şabloanelor au şi o formă fizică de reprezentare, constituindu-se în entităţi text. Dacă entităţile rezultate se încadrează în direcţia dată de şabloane, atunci ele se numesc entităţi bazate.

Analiza entităţilor bazate conduce la concluzii privind diferenţele dintre ele şi modalităţile de alegere şi aplicare a celor mai eficiente în managementul organizaţiilor.

Vocabular şi subvocabular

Se consideră alfabetul A = {a_1, a_2, ..., a_n}, unde *n* reprezintă numărul de simboluri, iar a_i este simbolul de pe poziția *i* din alfabet. Cu simbolurile alfabetului se construiesc cuvintele vocabularului V = {c_1, c_2, ..., c_m}, unde *m* reprezintă numărul de cuvinte din vocabular, iar c_j reprezintă cuvântul de pe poziția *j* din vocabular.

Dintre cuvintele vocabularului V, numai unele sunt reprezentative pentru un domeniu al ştiinţei, tehnicii, artei, economiei sau practicii de zi cu zi.

Se consideră domeniile D_1, D_2, …, D_{nD}, unde *nD* reprezintă numărul de domenii. Prin domeniu se înțelege un ansamblu de cunoştinţe pe baza cărora sunt derulate activități de o anumită natură, având caracteristici specifice. Caracteristicile domeniilor sunt, [IVAN05c]:

- termeni specifici; sunt descrise procese, instrumente, utilaje, activități, caracteristici de calitate a proceselor, produselor, serviciilor, stadii, etape, calificative;
- modalități de asociere a cuvintelor; se descriu procesele şi fenomenele care se manifestă într-un domeniu;
- termeni comuni; domeniile nu sunt complet separate unele de celelalte.

Prin reprezentativitatea vocabularului se evidențiază măsura în care cuvintele descriu, ilustrează, sugerează un obiect, fenomen, proces dintr-un domeniu.

Pentru domeniul D, se extrage un subvocabular SV_i = {c_{i1}, c_{i2}, ..., c_{ik}}, unde *k* reprezintă numărul de termeni specifici domeniului D, iar c_{ij} reprezintă cuvântul de pe poziția *j* din subvocabularul *i*.

După cum este alcătuit subvocabularul SV_i, rezultă $SV_i \subset V$.

Se consideră un text T constituit din cuvintele subvocabularului SV_x:

$$T = < c_{x1}\ c_{x2}\ \ldots\ c_{xH} >$$

Lungimea textului T măsurată în cuvinte este de H cuvinte.

Un text construit în sfera de cuprindere a domeniului D, trebuie să conțină un număr minim de termeni care aparţin domeniului pentru a fi considerat reprezentativ.

Textul T reprezintă un text de referinţă şi e utilizat la construirea de texte TT_1, TT_2, ..., TT_{nb}, unde *nb* reprezintă numărul de texte bazate.

În continuare, textele TT_i, i=1,2, ..., nb, se numesc texte bazate deoarece:

- se construiesc cu cuvinte din vocabularul V;
- este obligatorie folosirea unei ponderi însemnate din cuvintele subvocabularului SV_x.

Dacă se consideră vocabularul V, subvocabularul SV_x şi textul bazat TT_i, obiectivul de bază al unei analize este de a stabili măsura în care la generarea textului TT_i au fost întrebuinţate cuvintele subvocabularului SV_x.

În cazul în care cu cuvinte din subvocabularul SV se construieşte un text T, analiza trebuie adâncită la a stabili măsura în care textul bazat TT_i utilizează cuvinte din textul T construit pe subvocabularul SV.

Gradul de includere

Se consideră vocabularul:

V =< ortogonalitate metrici alfabet mulţime matrice indicatori software >

şi subvocabularul:

SV = < ortogonalitate metrici alfabet mulţime software >

Se consideră textele:

T = < se defineşte ortogonalitate mulţime precum şi ortogonalitate alfabet şi software, pe baza cărora rezultă gradul de asemănare între alfabet şi software >

şi:

TT_i = < rezultă indicatori asemănare alfabet şi software, precum şi matrice ortogonalitate alfabet şi software>

Se construieşte tabelul frecvenţelor de utilizare a cuvintelor din vocabularul V şi subvocabularul SV în textul T, respectiv în textul bazat TT_i, tabelul 1.

Frecvențele de utilizare a cuvintelor în T şi TT_i

Tabel 1

Vocabular	Subvocabular	T	TT_i
ortogonalitate	ortogonalitate	2	1
metrici	metrici	0	0
alfabet	alfabet	2	2
mulțime	mulțime	1	0
software	software	2	2
matrice	-	0	1
indicatori	-	0	1

Se calculează indicatorul grad de încredere G, conform expresiei analitice:

$$G = \frac{\sum_{j=1}^{nsv} \alpha_j}{\sum_{j=1}^{nsv} \beta_j}$$

unde:

α_j – variabilă booleană care ia una din valorile:
1, pentru utilizarea cuvântului c_{1j} în textul T şi în textul bazat TT_i;
0, când cuvântul c_{ij} lipseşte cel puțin dintr-un text, T sau TT_i;

β_j – variabilă booleană egală cu 1 atunci când cuvântul c_{ij} din subvocabularul SV este prezent în textul T;

nsv – numărul de cuvinte din subvocabularul SV.

Folosind datele din tabelul 1, se construieşte matricea variabilelor booleene α_j şi β_j, tabelul 2.

Modul de utilizare a cuvintelor din TT_i

Tabel 2

Vocabular	Subvocabular	T	TT_i	α_j	β_j
ortogonalitate	ortogonalitate	2	1	1	1
metrici	metrici	0	0	0	0
alfabet	alfabet	2	2	1	1
mulțime	mulțime	1	0	0	1
software	software	2	2	1	1
matrice	-	0	1	0	0
indicatori	-	0	1	0	0
Total	**-**	**8**	**7**	**3**	**4**

Gradul de încredere pentru datele din tabelul 2 este G = 0,75. Aceasta înseamnă că textul bazat TT_i acoperă într-o proporție de 75% domeniul specificat în entitatea T.

Se consideră textul:

TT_j = < rezultă indicatori de asemănare a simbolurilor, precum şi matrice de asemănare pentru simboluri >

Analiza textului TT_j conduce la frecvențele de apariție înregistrate în tabelul 3.

Frecvențele de utilizare a cuvintelor în T şi TT_j

Tabel 3

Vocabular	Subvocabular	T	TT_j
ortogonalitate	ortogonalitate	2	0
metrici	metrici	0	0
alfabet	alfabet	2	0
mulțime	mulțime	1	0
software	software	2	0
matrice	-	0	1
indicatori	-	0	1

Se construieşte tabelul variabilelor booleene α_j şi β_j.

Modul de utilizare a cuvintelor din TT_j

Tabel 4

Vocabular	Subvocabular	T	TT_j	α_j	β_j
ortogonalitate	ortogonalitate	2	0	0	1
metrici	metrici	0	0	0	0
alfabet	alfabet	2	0	0	1
mulțime	mulțime	1	0	0	1
software	software	2	0	0	1
matrice	-	0	1	0	0
indicatori	-	0	1	0	0
Total	**-**	**8**	**2**	**0**	**4**

Pentru valorile din tabelul 4, se calculează G = 0. Acesta înseamnă că textul TTj nu are nimic în comun cu T, nefiind utilizate cuvinte din subvocabularul SV.

Pentru textul:

TT_h = < rezultă o mulțime de indicatori de asemănare alfabet şi software, precum şi matrice ortogonalitate alfabet şi software>

se construieşte matricea:

Modul de utilizare a cuvintelor din TT_h

Tabel 5

Vocabular	Subvocabular	T	TT_h	α_j	β_j
ortogonalitate	ortogonalitate	2	1	1	1
metrici	metrici	0	0	0	0
alfabet	alfabet	2	2	1	1
mulțime	mulțime	1	1	1	1
software	software	2	2	1	1
matrice	-	0	1	0	0
indicatori	-	0	1	0	0
Total	**-**	**8**	**8**	**4**	**4**

Gradul de încredere pentru calculat pentru TT_h este G = 1. Textul bazat TT_h acoperă într-o proporţie de 100% domeniul specificat în entitatea T.

Gradul de externalizare evidenţiază măsura în care în textul TT_i sunt utilizate cuvinte din vocabularul V care nu aparţin subvocabularului, V \ SV.

Se numără cuvintele distincte care aparţin mulţimii V \ SV. Se consideră variabila booleană θ_j care ia una din valorile:

- 1, când cuvântul c_{ij} aparţine mulţimii V \ SV;
- 0, în caz contrar.

Gradul de externalizare G_e se determină conform relaţiei:

$$G_e = \frac{\sum_{i=1}^{nv} \theta_i}{\sum_{j=1}^{nsv} \alpha_j}$$

unde:

θ_j – variabilă booleană pentru apartenenţa cuvântului c_{ij} la mulţimea V \ SV;

α_j – variabilă booleană pentru apartenenţa cuvântului c_{ij} la textele T şi TT_i;

nsv – numărul de cuvinte al subvocabularului SV;

nv – numărul de cuvinte al vocabularului V.

Pentru valorile din tabelul 2, G_e = 0,66. În cea mai mare parte, cuvintele utilizate sunt din subvocabularul SV. Astfel, entitatea TT_i este una bazată.

Dacă G_e tinde către zero, atunci textul TT_i este construit folosind restricţiile impuse prin definirea subvocabularului SV.

Daca G_e tinde către 1 rezultă că textul TT_i operează cu cuvinte care aparţin în cea mai mare parte vocabularului V. Subvocabularul nu este bază în generarea de text, condiţie necesară ca textul TT_i să se numească text bazat.

Omogenitatea şi ortogonalitatea textelor bazate

Textele construite pe subvocabularul SV au ca obiectiv devenirea de repere în elaborarea de texte bazate.

Pentru elaborarea de referate asupra unei lucrări, se alcătuieşte un subvocabular din textul lucrării literare. Indicatorii G şi G_e cuantifică măsura în care a fost abordată opera literară, în raport cu modul în care autorii referatelor utilizează cuvintele subvocabularului.

Textele bazate sunt actualizate pentru modul în care utilizează cuvintele textului T, respectiv, cuvintele din subvocabularul SV.

Culegerile tipărite de analize literare presupun o mulțime de opere literare, T_1, T_2, ..., T_{nl} şi un subvocabular, unde *nl* reprezintă numărul de opere literare. Analizele literare propuse de autorii culegerilor tipărite sunt modele date sub forma textelor TT_1, TT_2, ..., TT_{nl}, nl = nb. Se formează perechile (T_1, TT_1), (T_2, TT_2), ..., (T_{nl}, TT_{nl}). Fiecărei opere îi corespunde un model de analiză literară. Autorii culegerii tipărite de analize literare urmăresc maximizarea nivelului indicatorului G şi minimizarea indicatorului G_e.

Cei care elaborează referate, lucrări de clasă, teze folosind un text de bază T_i produc textele TT_{i1}, TT_{i2}, ..., $TT_{i,nb}$. Prin calculul indicatorilor G şi G_e se măsoară gradul de utilizare a cuvintelor din subvocabularul SV, respectiv gradul de utilizare a altor cuvinte decât cele din subvocabular.

Textele bazate trebuie să fie omogene în raport cu modul în care referă cuvintele textului de bază.

Se calculează un indicator agregat al omogenității textelor bazate, $\overline{G}$:

$$\overline{G} = \sqrt[nb]{\prod_{i=1}^{nb} g_i}$$

unde g_i reprezintă valoarea indicatorului de apartenență G pentru TT_{ki}.

Gradul de omogenitate a entității TT_{kj} în raport cu apartenența la subvocabularul SV, se determină conform relației:

$$G_O = \frac{g_j - \overline{G}}{\overline{G}}$$

În acelaşi timp, textele TT_{i1}, TT_{i2}, ..., $TT_{i,nb}$ trebuie să fie ortogonale.

Pentru:

V =< ortogonalitate metrici alfabet mulțime matrice indicatori software >
SV = < ortogonalitate metrici alfabet mulțime >

se consideră textele bazate:

TT_i = < rezultă indicatori asemănare alfabet şi software, precum şi matrice ortogonalitate alfabet şi software>

TT_j = < se defineşte ortogonalitate mulțime precum şi ortogonalitate alfabet şi software>

Tabelul frecvențelor de utilizare a cuvintelor din vocabularul V şi subvocabularul SV este:

Frecvențele de utilizare a cuvintelor în TT_i şi TT_j

Tabel 6

Vocabular	Subvocabular	TT_i	TT_j
ortogonalitate	ortogonalitate	1	2
metrici	metrici	0	0
alfabet	alfabet	2	1
mulțime	mulțime	0	1
matrice	-	1	0
indicatori	-	1	0
software	-	2	1

Ortogonalitatea textelor bazate TTi şi TTj în raport cu utilizarea cuvintelor din vocabularul V se determină conform expresiei:

$$H(TT_i, TT_j) = \frac{\sum_{i=1}^{nv} \gamma_i}{nv}$$

unde:

γ_i – variabilă booleană ce ia una din valorile:

0, dacă C_k este utilizat în TT_i şi TT_j; de asemenea, pentru cazul în care c_k nu apare nici în TT_i, nici în TT_j;

1, în caz contrar;

nv – numărul de cuvinte din V.

Pentru valorile din tabelul 6, $H(TT_i, TT_j) = 0,42$.

Dacă $H(TT_i, TT_j)$ tinde spre 1 înseamnă că textele bazate TT_i şi TT_j sunt ortogonale. Dacă $H(TT_i, TT_j)$ tinde spre zero, atunci entitățile analizate sunt aproape identice.

Pentru exemplul considerat, diferențele nu sunt semnificative. Se afirmă că textele bazate TT_i şi TT_j fac parte din acelaşi domeniu.

Software pentru analiza textelor bazate

Există un tezaur de fizică:

TZ = < optica, foton, ondulatoriu, energetic, unde, fotometria, electromotor, rezonanța, lumina, polarizare, dispersia, radiolocația, sunetul, osciloscop, rezistor, circuit, bobina, condensator, curent, oscilație, principiu, propagare >

Din cuvintele tezaurului de fizică se extrag termenii pentru o programă analitică.

Termenii formează subvocabularul:

SV = < optica, ondulatoriu, energetic, radiolocația, sunetul, rezistor, principiu, bobina, condensator, electromotor, rezonanta, circuit >

iar programa analitică este dată de:

T = < optica ondulatoriu principiu radiolocația circuit >

Autorii scriu manualele TT_1, TT_2, TT_3 şi acestea sunt texte bazate pentru că trebuie să aibă cuvintele din programă.

TT_1 =< circuit dispersie unde radiolocația >
TT_2 = < principiu unde circuit bobina lumina rezistor foton >
TT_3 = < optica propagare circuit ondulatoriu radiolocația bobina circuit condensator curent >

Aplicația EBASOF are o structură modularizată, vezi figura 1.

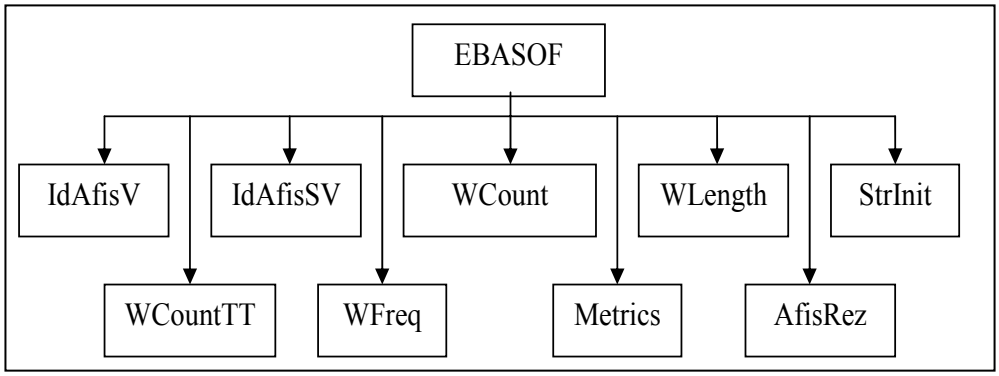

Figura 1 Structura aplicației EBASOF

Determinarea indicatorilor privind gradele de încredere şi de externalizare se realizează prin intermediul aplicației EBASOF, elaborată în limbajul C++. Aplicația cuprinde următoarele module:

IdAfisV realizează identificarea şi afişarea cuvintelor din vocabularul V; constă în deschiderea fişierului care conține cuvintele vocabularului, parcurgerea fişierului pentru identificarea şi afişarea cuvintelor; numele fişierului este furnizat de la tastatură;

IdAfisSV efectuează identificarea şi afişarea cuvintelor dintr-un subvocabular SV_i; presupune deschiderea fişierului care conține cuvintele din subvocabular şi traversarea sa pentru identificarea şi afişarea cuvintelor; numele fişierului se introduce de utilizator de la tastatură;

WCount realizează determinarea numărului de cuvinte din vocabular şi subvocabular; constă în identificarea cuvintelor din vocabular şi subvocabular, însoțită de contorizare; această se realizează prin identificarea şi numărarea simbolurilor separator din şirul de caractere care formează entitatea subvocabular;

WLength implementează determinarea lungimilor cuvintelor din vocabular şi subvocabular; lungimea cuvintelor se determină ca număr de caractere care sunt utilizate în construirea acestora;

StrInit permite iniţializarea structurilor şi verificarea apartenenţei subvocabularului SV la vocabularul V; structurile utilizate pentru iniţializare sunt de masive unidimensionale şi bidimensionale; verificarea este necesară pentru a asigura calitate ridicată pentru analiza textelor bazate;

WCountTT realizează determinarea numărului de cuvinte din SV_i utilizate în textele TT_i; cuvintele subvocabularului SV_i sunt identificate în textele TTi, date de utilizator de la tastatură; aceste frecvenţe de apariţie sunt utilizate în calculul metricilor de apartenenţă;

WFreq efectuează determinarea şi afişarea frecvenţelor de apariţie a cuvintelor din V şi SV_i incluse în textele TT_i; se determină numărul de cuvinte din V care apar în textele TT_i; se afişează frecvenţele de apariţie ale cuvintelor din V şi SV_i sub formă tabelară;

Metrics realizează implementarea modelelor asociate metricilor; presupune pregătirea datelor de intrare prin considerarea elementelor structurale ale metricilor de includere şi externalizare; modelele verifică apartenenţa cuvintelor din TT_i care se regăsesc în V şi SV_i în raport cu cuvintele care se regăsesc numai în subvocabular sau care sunt în V şi nu aparţin de SV_i;

AfisRez efectuează afişarea rezultatelor; sunt furnizate valorile obţinute prin aplicarea modelelor asociate metricilor pentru tabelele cu frecvenţele de apariţie construite.
Execuţia aplicaţiei EBASOF permite obţinere gradelor de încredere:

$G(TT_1) = 0,40$
$G(TT_2) = 0,40$
$G(TT_3) = 0,80$

şi gradele de externalizare:

$Ge(TT_1) = 0,29$
$Ge(TT_2) = 0,25$
$Ge(TT_3) = 0,57$

pentru tezaurul de fizică T şi textele bazate TT_1, TT_2 şi TT_3.

Valorile indicatorului G sunt relativ ridicate. Interpretarea lor se rezumă la faptul că mare parte din cuvintele textelor TT_i, i = 1, 2, 3 fac parte din textul T.

Pentru valorile indicatorului G_e, cuvintele utilizate sunt din subvocabularul SV. Entitățile TT_i, i = 1, 2, 3 sunt entități bazate.

Concluzii

Analiza textelor bazate este foarte importantă pentru determinarea asemănării a două texte având la baza acelaşi vocabular de bază sau acelaşi subvocabular.

Calculând gradul de încredere şi gradul de externalizare al textelor se verifică gradul de proveniență al unui text dintr-un vocabular sau dintr-un alfabet.

Lucrările de specialitate comportă o mare utilizare a software-ului de analiză al textelor bazate, întrucât rezultatele furnizate de acesta probează apartenența lucrărilor la un domeniu de specialitate.

Bibliografie

1. BALLOU, D. TAYI, G. K. „Enhancing Data Quality in Data Warehouse Environments”, *Communications of the ACM*, vol. 42, nr. 1, 1999, pp. 73 – 78
2. IVAN, I. POPA, M. *Entități text – dezvoltare, analiză, evaluare*, Bucureşti, Editura ASE, 2005
3. IVAN, I. POPA, M. TOMOZEI, C. „Reingineria entităților text”, în *Revista Română de Informatică şi Automatică*, vol. 15, nr. 2, 2005, pp. 15 – 28

4.	IVAN, I. POPA, M.	*Text Entities Reprezentativeness*, Conferința Internațională „The Impact of European Integration on the National Economy", Babeş-Bolyai University, Cluj-Napoca, October 28 – 29, 2005
5.	IVAN, I. POPA, M. BOJA, C.	*Operații pe entități text*, Conferința Ştiințifică Internațională „Binomul sărăcie-bogăție şi integrarea României în Uniunea Europeană", Sibiu, 20 – 21 mai 2005, vol. 3, pp. 464 – 473
6.	IVAN, I. POPA, M. CAPISIZU, S. BREDA, L. FLORESCU, B.	*Clonarea informatică*, Bucureşti, Editura ASE, 2003
7.	POPA, M.S	„Text Entities Metrics", în *Informatica Economică*, vol. 9, nr. 2, 2005, pp. 56 – 60
8.	POPA, M.	*Structured Text Entities Metrics*, The Proceedings of „The 36th Informational Scientific Symposium of METRA", Bucureşti, 26 - 27 mai 2005, pp. 486 – 491
9.	POPA, M.	*Text Entities Topographic Characteristic*, Proceedings of the Seventh International Conference on Informatics in Economy „Information & Knowledge Age", Bucureşti, 19 – 20 mai 2005, pp. 1115 – 1121
10.	SMEUREANU, I. DÂRDALĂ, M.	*Programarea orientată obiect în limbajul C++*, Bucureşti, Editura CISON, 2002
11.	***	www.research.ibm.com