Uniformity – quality characteristic of text entities

Prof.dr. Ion IVAN, asist. Marius POPA
Catedra de Informatică Economică, ASE Bucureşti

In this paper there are presented some aspects concerning the concepts about the text entities and some criteria for classification of these ones. It is defined the concept of uniformity as a quality characteristic of the text entities. It is presented the text entity with linear structure and how this entity type is built. Also, a quantitative analyze is offered together with metrics concerning the uniformity measuring.
Keywords: *uniformity, text entity, metrics.*

1 Tipologii de entităţi text

Există numeroase criterii de clasificare a entităţilor text. După criteriul dimensiunii sau lungimii, entităţile text sunt entităţi de lungime mică, entităţi de lungime medie şi entităţi de lungime mare. Apartenenţa la una dintre grupe se realizează strict prin referire la elementele unei colectivităţi bine delimitate. Dacă se consideră o mulţime de proiecte se stabileşte pentru fiecare element lungimea, se identifică proiectul de lungime minimă şi proiectul de lungime maximă, se stabileşte lungimea medie folosind indicatorul de medie aritmetică simplă şi se procedează la gruparea proiectelor. Pentru alte entităţi se procedează în acelaşi fel. Există şi alte modalităţi de a obţine grupe omogene după criteriul lungimii.

În funcţie de complexitate, entităţile text sunt cu nivel de complexitate redus, cu nivel de complexitate mediu şi, respectiv, nivel maxim de complexitate.

După criteriul omogenităţii, entităţile sunt omogene, respectiv entităţi neomogene.

Uniformitatea este o caracteristică de calitate importantă care vizează atât nivelurile structurii unei entităţi cât şi elementele ce alcătuiesc structura. O entitate este uniformă dacă elementele ce intră în componenţa structurii au acelaşi nivel de omogenitate, au acelaşi nivel de complexitate şi lungimile nu diferă semnificativ de la o componentă la alta.

Uniformitatea componentelor, uniformitatea dispunerii acestora în structura entităţii influenţează din punct de vedere calitativ întreaga entitate, contribuind la obţinerea unei construcţii echilibrată în ansamblul ei.

O entitate construită în spiritul uniformităţii presupune stabilirea unui obiectiv clar, identificarea resurselor, definirea de activităţi, alegerea de instrumente, culegerea de date, definirea de algoritmi de prelucrare, obţinerea de rezultate care converg spre atingerea obiectivului. Orice entitate text înzestrată cu caracteristica de uniformitate presupune existenţa unui vocabular unic, referirea elementelor din vocabular cu înregistrarea frecvenţelor de referire, astfel încât distribuţia pe care aceste frecvenţe o urmează este, de asemenea, o distribuţie uniformă. Tratamentul entităţilor text pentru evidenţierea caracteristicii de uniformitate presupune strict utilizarea tehnicilor şi metodelor statistice.

2. Entităţi text cu structură liniară

Se consideră o entitate text ET formată din blocurile B_1, B_2, …, B_n. Într-o primă fază blocurile se consideră uniforme dacă textele pe care le includ sunt realizate folosind cuvintele vocabularului V. În entitatea liniară se identifică:

B_1 – blocul iniţial, caracterizat prin legătură într-un singur sens spre unul dintre blocurile intermediare;

B_i – bloc intermediar, caracterizat prin legătură cu blocul precedent B_{i-1} şi legătură cu blocul următor B_{i+1};

B_n – blocul final, caracterizat printr-o singură legătură cu blocul precedent B_{n-1}, figura 1.

Uniformitatea structurii liniare a unei entităţi text, se evidenţiază, prin nivelul de consistenţă care se înregistrează între blocurile adiacente. Enunţurile din blocul B_i se dezvoltă în blocul B_{i+1} şi se bazează pe elemente formulate deja în blocul B_{i-1}.

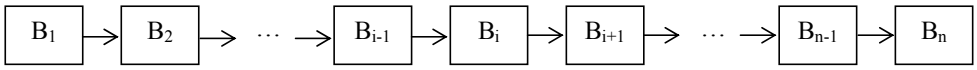

Fig. 1. Structura liniară a entității ET

La nivelul fiecărui bloc al entității text, se înregistrează şiruri de cuvinte ale vocabularului V, care nu trebuie să se repete în blocurile text următoare, consistența fiind asigurată de caracterul progresiv al dezvoltării textelor în blocuri. Dacă în blocul B_i este dată o definiție pentru un anumit concept, în blocurile B_j, $j > i$, se evită prezentarea altor definiții pentru respectivul concept chiar dacă esența rămâne aceeaşi. Cu atât mai mult este necesar ca în blocul B_j să nu apară o altă definiție care vine în contradicție cu definiția dată în blocul B_i. În acelaşi fel, se asigură consistență şi în cazul modelelor şi în cazul prezentării de algoritmi. Pentru modele se evită construirea mai multor ecuații destinate stabilirii legăturilor dintre o aceeaşi variabilă endogenă şi diferite variabile exogene. Pentru algoritmi se evită definirea de paşi care au rolul de a anula efectele prelucrărilor pe care le presupun paşii precedenți. Consistența este însoțită de un echilibru ce trebuie asigurat nivelului de detaliere. Uniformitatea entității text presupune ca la nivelul fiecărui bloc detalierea să se facă pe un număr de niveluri ce nu diferă semnificativ de numărul de niveluri din blocurile adiacente.

Dacă se doreşte realizarea unei entități text de lungime 10.000 cuvinte pentru prezentarea istoriei unui popor, urmărindu-se uniformitatea nivelului de abordare, entitatea text se structurează astfel: 5% - introducere, 5% - concluzii, 15% - perioada contemporană, 15% - perioada modernă, 15% - perioada preindustrială, 15% - evul mediu, 15% - antichitate, 15% - preistorie. Rezultă că autorii trebuie să se concentreze în aşa fel încât, folosind un vocabular V în mod uniform, pe un text de aproximativ 1.500 de cuvinte să redea evenimente, personalități, caracteristici ale fiecărei perioade.

Pentru fiecare perioadă se construieşte câte o structură arborescentă cu acelaşi număr de niveluri, frunzelor structurii corespunzându-le texte de lungime medie obținută prin împărțirea lungimii blocului la număr de frunze ale arborescenței. Orice altă abordare produce dezechilibre care afectează uniformitatea în interiorul componentelor, aspect ce se propagă la nivelul întregii entități.

În cazul în care se elaborează un proiect pentru obținerea finanțării, asigurarea credibilității în procesul de evaluare este dată de menținerea unui nivel ridicat de uniformitate a descrierilor pentru activități, resurse, termene, interacțiuni, riscuri. Uniformitatea vizează acelaşi nivel ridicat de detaliere care evidențiază faptul că echipa care implementează proiectul cunoaşte foarte bine domeniul şi are experiența necesară.

În cazul dezvoltării activităților artistice, ca de exemplu scrierea unei poezii, menținerea ritmului şi rimelor este condiție pentru asigurarea uniformității entității text.

La soluționarea unei probleme în plan informatic prin scrierea de software în limbaje de programare, uniformitatea este rezultatul unei decizii impusă de complexitatea problemei. Consecința directă este alegerea limbajului de programare, resurselor disponibile în biblioteci şi a tehnicilor şi metodelor de programare pe care aplicația le impune în vederea atingerii obiectivului stabilit.

3. Analiza cantitativă

Există trei moduri de a abordare a analizei cantitative. Primul mod vizează existența vocabularului V specific domeniului pentru care se construieşte entitatea ET. Toate analizele vizează modul în care blocurile entității ET referă cuvintele vocabularului V. Analizoarele de text, în acest caz, au menirea de a construi matrice ale frecvențelor de referire pentru cuvinte în fiecare bloc conform tabelului 1.

Tabelul 1. Matricea frecvențelor de referire

Cuvânt din vocabularul V	**Blocuri**					
	B_1	B_2	…	B_j	…	B_n
c_1	f_{11}	f_{12}	…	f_{1j}	…	f_{1n}
c_2	f_{21}	f_{22}	…	f_{2j}	…	f_{2n}
c_3	f_{31}	f_{32}	…	f_{3j}	…	f_{3n}
…						
c_i	…	…	…	f_{ij}	…	…
…						
c_m	f_{m1}	f_{m2}	…	f_{mj}	…	f_{mn}

În tabelul 1 f_{ij} reprezintă frecvența de apariție a cuvântului c_i în blocul B_j. Analiza cantitativă la nivelul utilizării în blocuri a cuvintelor din vocabular se identifică ortogonalitatea perfectă dacă în blocul B_j sunt folosite cuvintele $c_{j1}, c_{j2}, \ldots, c_{jn}$. Dacă pentru blocurile B_{j-1}, B_j şi B_{j+1} se identifică elementele de complementaritate, tabelul 2, blocurile înregistrează nivelul maxim de ortogonalitate.

Tabelul 2. Nivelul maxim de ortogonalitate a blocurilor

Cuvânt din vocabularul V	**Blocuri**					
	…	B_{j-1}	B_j	B_{j+1}	…	…
…	…			…	…	…
$c_{i-1,1}$	…	$f_{i-1,1}$			…	…
$c_{i-1,2}$	…	$f_{i-1,2}$			…	…
$c_{i-2,3}$	…	$f_{i-1,3}$			…	…
…	…	…				…
$c_{i-1,k}$	…	$f_{i-1,k}$			…	…
$c_{i,1}$	…		$f_{i,1}$			…
$c_{i,2}$	…		$f_{i,2}$		…	…
$c_{i,3}$	…		$f_{i,3}$		…	…
…	…		…		…	…
$c_{i,l}$	…		$f_{i,l}$		…	…
$c_{i+1,1}$	…			$f_{i+1,1}$	…	…
$c_{i+1,2}$	…			$f_{i+1,2}$	…	…
$c_{i+1,3}$	…			$f_{i+1,3}$	…	…
…	…			…	…	…
$c_{i,m}$	…			$f_{i+1,m}$	…	…
…	…				…	…

Complementaritatea evidențiază faptul că entitatea este construită din componente relativ independente.

Uniformitatea este dată de faptul că împrăştierea față de medie a frecvențelor de folosire a cuvintelor din vocabulare nu diferă semnificativ de la un bloc la altul. Testele statistice de egalitate medii şi dispersii au menirea de a verifica ipotezele de uniformitate a componentelor entității.

Algoritmul pentru evaluarea uniformității unei entități cu blocuri complementare presupune următorii paşi:

- verificarea complementarității;
- calculul indicatorului de medie a frecvenței cuvintelor din fiecare bloc;
- calculul dispersiilor;
- aplicarea testului de egalitate medii;
- aplicarea testului de aplicare dispersii;
- concluzionarea asupra uniformității blocurilor dacă ipotezele privind egalitatea mediilor şi dispersiilor se verifică.

Al doilea mod de analiză a uniformității în cadrul entității text este bazat pe construirea vocabularului pe măsură ce se traversează textele din blocurile ce compun entitatea.

Uniformitatea în acest caz vizează atât frecvențele de folosire a cuvintelor, cât şi modul în care se realizează evidențierea laturilor comune ale blocurilor, figura 2.

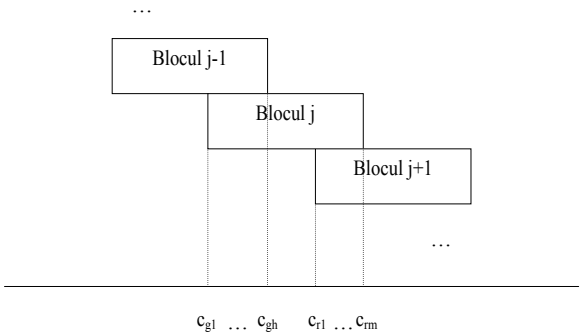

Fig. 2. Laturile comune ale blocurilor

Blocurile B_{j-1} şi B_j au în comun cuvintele c_{g1}, c_{g2}, …, c_{gh}, iar blocurile B_j şi B_{j+1} au în comun cuvintele c_{r1}, c_{r2}, …, c_{rm}. Aceste subvocabulare comune permit efectuarea tranziției de la un bloc la celălalt în cadrul structurii liniare. Al treilea mod de analiză a uniformității vizează o combinație între primele două astfel încât, pornind de la vocabularul V, se construieşte tabelul de frecvențe şi se elimină cuvintele cu frecvențe nule, după care se realizează o reordonare descrescătoare în cadrul fiecărui bloc după totalul de frecvențe al fiecărui cuvânt. Se identifică uniformitatea cu care se referă din aproape în aproape cuvintele în cadrul blocurilor.

Analiza cantitativă se continuă cu studiul corelațiilor dintre frecvențele cuvintelor care apar în textele uniforme din blocuri. În cazul în care coeficienții de corelație au valori în jurul lui zero, corelația este considerată ca fiind slabă, iar ca primă consecință entitatea text are un nivel de uniformitate redus. În cazul în care coeficientul de corelație tinde spre valoarea 1, uniformitatea este ridicată şi are un caracter ascendent, creşterii frecvenței de utilizare a unor cuvinte ce presupun adâncirea detalierii într-un bloc îi va corespunde, de

asemenea, creşterii frecvenţei de folosire a altor cuvinte, fapt care menţine nivelul de detaliere şi în blocul adiacent. În cazul în care coeficientul de corelaţie tinde către valoare -1, nivelul de detaliere se menţine între blocuri, dar abordarea are un caracter simetric, adică într-un bloc se porneşte de la detalii specifice proceselor de analiză către sinteză, iar în blocul adiacent se porneşte de la elementele sintetice spre detalii.

4. Construirea unei entităţi text cu structură liniară

Se consideră că este necesar să se construiască o entitate text cu structură liniară ETL. Această entitate text este formată din textele T_1, T_2, T_3, T_4, T_5, T_6, T_7, T_8, T_9 şi T_{10}, având lungimile L_1, L_2, L_3, L_4, L_5, L_6, L_7, L_8, L_9, respectiv, L_{10}.

Prin traversarea textului TT rezultat din reuniunea textelor T_1, T_2, T_3, T_4, T_5, T_6, T_7, T_8, T_9 şi T_{10}, se construieşte vocabularul acestei entităţi V_{ETL}.

Construirea acestei entităţi vizează ca blocul B_i asociat textului uniform T_i să corespundă structurii din figura 3. Frecvenţele de utilizare a cuvintelor vocabularului V_{ETL} în blocuri sunt reprezentate în tabelul a cărui formă grafică este dată în figura 4.

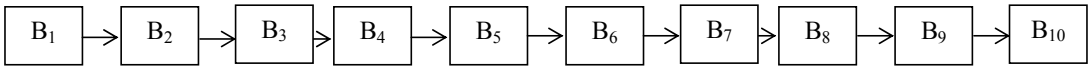

Fig. 3. Structura entităţii liniare ETL

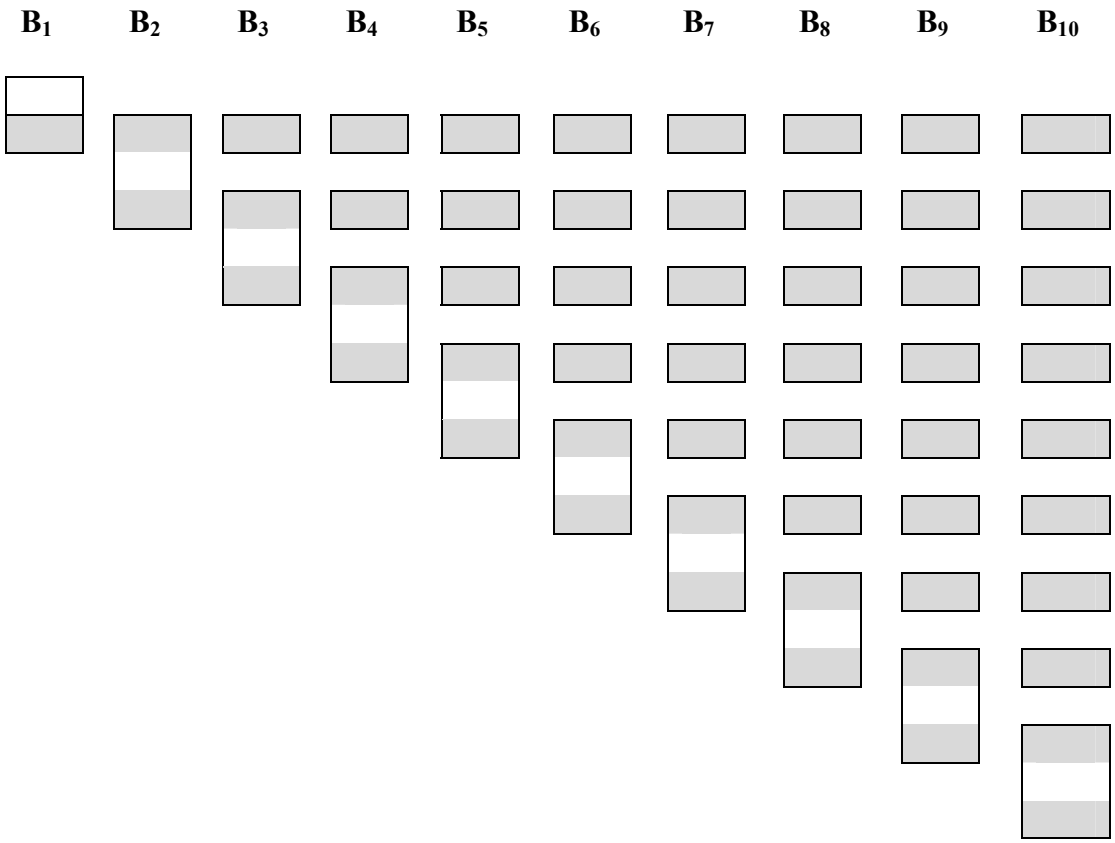

Fig. 4. Frecvenţele de utilizare a cuvintelor vocabularului V_{ETL}

Reiese din figura 4 uniformitatea referirii de cuvinte ale subvocabularului $V_{ETL}(B_1)$ în toate blocurile B_2, B_3, B_4, B_5, B_6, B_7, B_8, B_9 şi B_{10}. Progresiv, există elemente comune ale subvocabularului $V_{ETL}(B_i)$ în toate blocurile B_{i+1}, B_{i+2}, …, B_{10}. Un exemplu de construire a unei entităţi text liniare de acest tip îl reprezintă elaborarea unui articol care prezintă rezultatul unei cercetări.

T_1 – titlul articolului;

T_2 – rezumatul;

T_3 – definirea conceptelor de bază;

T_4 – prezentarea rezultatelor existente în literatura de specialitate;
T_5 – definirea problemei a cărei soluție îşi propune să o prezinte articolul;
T_6 – prezentarea soluției: model, algoritm, proces, tehnologie, metodă;
T_7 – analiza comparată cu ceea ce există deja;
T_8 – interpretarea rezultatelor;
T_9 – concluzii;
T_{10} – bibliografie.
În cele mai multe cazuri, se preferă dezvoltarea de entități text neliniare care au de regulă caracter neuniform, iar complexitatea ridicată impun efort suplimentar de traversare şi înțelegere. Există procedee mecanice de transformare a entităților neliniare în entități liniare, însoțite de creşterea nivelului de redundanță.

5. Indicatori ai uniformității

Uniformitatea este opusul diversității. Dacă se consideră indicatorul:

$$G = \frac{\sum_{i=1}^{k} h_i \log_2 h_i}{(\sum_{i=1}^{k} h_i)\log_2(\sum_{i=1}^{k} h_i)}$$ unde k – numărul de blocuri care definesc entitatea text; h_i – numărul de cuvinte diferite care apar numai în blocul B_i; se obține o măsură a uniformității unei entități text. Se consideră entitatea text definită ca în tabelul 3.
Se calculează gradul de omogenitate şi rezultă G = 0,5, ceea ce înseamnă că entitatea text are un nivel mediu de uniformitate.
Dacă G tinde către zero, uniformitatea este redusă, iar dacă G tinde către 1 uniformitatea se apropie de nivelul maxim.

6. Concluzii

Pentru studiul caracteristicii de uniformitate trebuie definiți indicatori care să extindă analiza pentru entități dezvoltate ca reuniuni de vocabulare. Este necesară automatizarea procesului de analiză a structurilor de blocuri şi de reorganizare a cuvintelor din vocabular pentru a se obține acele interdependențe ce evidențiază corelațiile directe, inverse sau chiar absența acestora şi care permit concluzii asupra nivelului de uniformitate din entitatea text.

Tabelul 3. Frecvențe de apariție a cuvintelor în blocuri

Cuvânt	Blocuri			
	B_1	B_2	B_3	TOTAL
c_0	4			**4**
c_1	3			**3**
c_2	4			**4**
c_3	7			**7**
c_4	2	3		**5**
c_5	4	1		**5**
c_6		4		**4**
c_7		4		**4**
c_8		2	3	**5**
c_9		4	2	**6**
c_{10}			5	**5**
c_{11}			2	**2**
TOTAL	**24**	**18**	**12**	**54**
MEDIA	**4,00**	**3,00**	**3,00**	

Numărul de cuvinte care nu sunt comune blocurilor este dat în tabelul 5.

Tabelul 5. Numărul de cuvinte specifice blocurilor

Indicator de frecvențe	Frecvențe de apariție	Blocuri de cuvinte nespecifice
h_1	4	B2, B3
h_2	2	B1, B3
h_3	2	B1, B2

În cazul unor structuri deosebit de complexe, metodele statistice de analiză dispersională şi metodele de analiză datelor sunt instrumente eficiente pentru evidențierea caracteristicii de uniformitate din entitatea text, dacă aceasta există.

Bibliografie

[IASI03] Alexandru Isaic-Maniu, Constantin Mitruț, Vergil Voineagu – *Statistică*, Editura Universitară, Bucureşti, 2003
[IASI95] Alexandru Isaic-Maniu, Constantin Mitruț, Vergil Voineagu – *Statistica pentru managementul afacerilor*, Editura Economică, Bucureşti, 1999
[INFO04] Sistem de evaluare a entităților bazate pe text - *Entități text şi caracteristici de calitate*, programul INFOSOC, raport de cercetare
[MARI84] Ion Marinescu – *Analiza factorială*, Editura Ştiințifică şi Enciclopedică, Bucureşti, 1984
[VADU70] Ion Văduva – *Analiza dispersională*, Editura Tehnică, Bucureşti, 1970