Software pentru evaluarea amprentei textului utilizând numere pseudoaleatoare

Prof.dr. Ion IVAN, prep. Marius POPA, prep. Adrian POCOVNICU
Catedra de Informatica Economica, A.S.E. Bucuresti

The print is aggregated information that is extracted from a text. It is presented a software product that implements an algorithm for text print construction. The random number generators are used to achieve the sample that is analyzed.
Keywords: *text print, sample, random numbers.*

Construirea esantionului de texte

În [POCO03] este definita amprenta unui text ca fiind un sir de valori care identifica în mod unic acel text. Se considera o multime de texte T_1, T_2, ..., T_n din care se extrage un esantion. Procedurile P_1, P_2, ..., P_k pentru generarea numerelor pseudoaleatoare care urmeaza legile de distributie D_1, D_2, ..., D_k conduc la obtinerea unor rezumate de texte T_1', T_2', …, T_n'. Indicatorii construiti utilizeaza parti din esantioanele de texte folosind sirurile de numere pseudoaleatoare generate. Se calculeaza amprenta agregata a textului. Se stabilesc modalitati de utilizare a amprentei agregate pentru identificarea clonelor din multimea initiala de texte.

Pentru construirea esantionului se pot folosi mai multe procedee:

- procedee de esantionare aleatoare: loterie, numere aleatoare, mecanic;
- esantionare dirijata;
- esantionare mixta.

Pentru determinarea volumului esantionului se efectueaza urmatoarele operatii:

- se estimeaza dispersia multimii din care se face esantionarea;
- se stabileste probabilitatea cu care esantionul va fi reprezentativ pentru multime initiala;
- i se asociaza probabilitatii stabilite valoarea tabelara a variabilei *Student*;
- se stabileste eroarea maxim admisa.

Volumul esantionului va fi $L_E = \frac{t^2_{n-1,a} \cdot s^2}{d^2}$

unde:

- $t^2_{n-1,a}$ este valoarea tabelara a variabilei *Student* cu n-1 grade de libertate;
- s^2 este dispersia de sondaj, care aproximeaza dispersia necunoscuta;
- d^2 este eroarea maxim admisa.

Procedeele de esantionare se aplica dupa determinarea volumului optim al esantionului, care asigura estimarea corecta a parametrilor multimii initiale.

Module pentru generare de numere pseudoaleatoare

Se considera un algoritm A_i pentru generarea numerelor pseudoaleatoare cu legea de distributie D_i, i = 1, 2, …, k. Algoritmul A_i presupune o valoare de start X_{0i} si un interval definit prin $[L^i_{inf}; L^i_{sup}]$, iar dupa o serie de calcule se obtine un numar $y_i \in [L^i_{inf}; L^i_{sup}]$ pseudoaleator, generat, pe care algoritmul îl garanteaza ca are o probabilitate de aparitie care urmeaza legea D_i. Algoritmul A_i este construit si supus unor teste si numai dupa ce este demonstrat ca genereaza numere pseudoaleatoare cu legea de distributie D_i, este folosit în aplicatii practice.

Recursivitatea algoritmului permite generarea unor siruri de numere pseudoaleatoare y_1, y_2, ..., y_n. Limbajul C++ contine functia *int rand ( )* pentru generarea numerelor pseudoaleatoare uniform distribuite. Procedura *PseudoRand* este utilizata pentru initializarea unui vector cu n componente cu numere pseudoaleatoare urmeaza o lege de distributie diferita de cea uniform distribuita. Codul sursa al procedurii este prezentat la adresa www.ivan.ase.ro/pseudo . Experimental, în aceasta procedura se observa ca functia de generare a numerelor pseudoaleatoare este periodica, cu o perioada mai mare de 150. În practica este optim a folosi algoritmi de ge-

nerare de numere pseudoaleatoare cu perioada infinita. În [***70] sunt prezentate proceduri pentru generare de numere pseudoaletoare cu legi de repartitie uniforme, normale, Gauss, Beta, X^2. Pentru generarea de numere pseudoaleatoare este utilizata repartitia Gauss.

Mecanismele de obtinere a sirurilor de numere pseudoaleatoare apartinând unui interval consta în împartirea intervalului initial într-un numar de subintervale egal cu numarul de valori ce trebuie generate. Se realizeaza punerea în corespondenta a subintervalului caruia apartine numarul pseudoaleator generat cu valoarea din multimea de baza formata din k elemente b_1, b_2, ..., b_k si calculeaza:

$$r = \frac{L_{sup} - L_{inf}}{k}$$

Punerea în corespondenta a subintervalelor cu valorile multimii de baza este data de tabelul 1.

Tabelul 1. Corespondenta subinterval – valoare

Subinterval	Valoare de baza
$[L_{inf}; L_{inf} + r)$	b_1
$[L_{inf} + r; L_{inf} + 2 \cdot r)$	b_2
…	…
$[L_{inf} + (k-1) \cdot r; L_{sup}]$	b_b

Daca la repetarea procesului de generare cu ordinul i se obtine $x_0 \in [L_{inf} + j \cdot r ; L_{inf} + (j+1) \cdot r)$, atunci $x[i] = b_{j+1}$. În cazul în care procedura realizeaza generarea de numere pseudoaleatoare în intervalul $[L_{inf} ; L_{sup}]$, iar problema necesita lucrul cu numere pseudoaleatoare în intervalul [A ; B] un mod de a realiza translatia pe noul interval este:

$x_0 \in [L_{inf} ; L_{sup}]$

$y_0 = A + (B-A) \cdot (x_0 - L_{inf}) / (L_{sup} - L_{inf})$

$y_0 \in [A ; B]$

Procedura pentru implementarea translatiei este prezentata la adresa www.ivan.ase.ro/pseudo .

În cadrul procesului de dezvoltare de aplicatii complexe, pentru obtinerea de amprente folosind numere pseudoaleatoare este necesara dezvoltarea unei clase saturate cu metode care sa acopere atât generarile cât si translatiile pentru cât mai multe legi de distributie.

Procedeul mecanic de esantionare este un procedeu cvasialeator. Operatia de alcatuire a esantionului în acest caz este precedata de stabilirea pasului de numarare. Acesta trebuie sa fie un numar întreg calculat ca raport între lungimea textului si lungimea esantionului.

Construirea esantionului presupune efectuarea urmatoarelor etape: se selecteaza la întâmplare un numar din prima grupa; se adauga succesiv pasul de numarare pâna la obtinerea celor L_E elemente. Asadar, fie p pasul de numarare, se alege aleator un numar cuprins între 1 si p. Se includ apoi în esantion unitatile de pe pozitiile: k + p, k+2p, k+3p, etc., pâna la epuizarea listei.

Modulul pentru extragerea esantioanelor din texte

Fisierele text sunt de regula de dimensiuni foarte mari, iar problemele de analiza sunt definite pe multimi de astfel de fisiere. La nivelul textului se efectueaza analize folosind esantioane. Procedura de extragere a esantionului din text este prezentata la adresa www.ivan.ase.ro/pseudo .

Vectorul x contine pozitiile din textul T ale caracterelor ce vor fi extrase. T[0] este caracterul de pe prima pozitie din textul T, T[1] cel de pe a doua pozitie s.a.m.d. Se construieste astfel un rezumat y cu caracterele extrase y[0] = T[x[0]], y[1] = T[x[1]], …, y[L_E-1] = T[x[L_E-1]]. Procedeul este prezentat în figura 1.

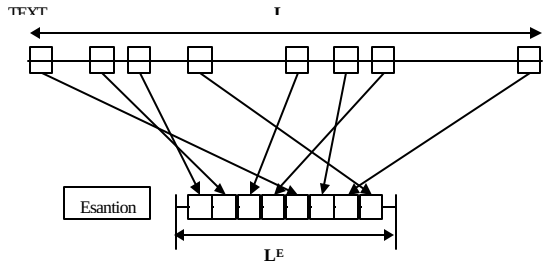

Fig. 1 Esantionarea unui text

Se stabileste alfabetul textului y si frecventa de aparitie a simbolurilor. De exemplu se considera textul T care are lungimea de 50 de cuvinte. Se presupune ca esantionul trebuie sa aiba L_E = 20 cuvinte. Dupa procedura de

generare se obtin 20 de numere în intervalul [0;49], unde lungimea textului T este 49. Textul y va avea lungimea egala cu L_E si continutul dat de pozitiile generate. Se calculeaza media, dispersia frecventelor de aparitie a cuvintelor, coeficientul de variatie, precum si numarul de inversiuni pentru a ordona frecventele simbolurilor alfabetului esantionului din textul T si frecventele cuvintelor din vocabularul textului esantionului. În acest fel s-au obtinut patru indicatori de baza ai esantioanelor obtinute prin generare de numere pseudoaleatoare.

Modulul pentru construirea amprentei esantionului

Amprenta este un sir de valori care definesc un text. Daca se extrage un esantion reprezentativ pentru care se construieste amprenta, atunci aceasta este, de asemenea, reprezentativa si pentru text. Se propune realizarea unui indicator agregat. În intervalul [2 ; 10] se genereaza 4 numere pseudoaleatoare uniform distribuite care se dispun în ordine descrescatoare $a_1>a_2>a_3>a_4$. Se considera x_1, x_2, x_3, x_4 valorile celor patru indicatori:

- coeficientii de variatie ai alfabetului si vocabularului esantionului (cu valori pozitive subunitare);
- numarul inversiunilor necesar pentru sortarea frecventelor simbolurilor din alfabetul esantionului si pentru sortarea frecventelor cuvintelor din vocabularul esantionului (valoare pozitiva subunitara).

Indicatorul G este un numar care se numeste amprenta agregata:

$$G = \frac{a_1^{x_1} \cdot a_2^{x_2} \cdot a_3^{x_3} \cdot a_4^{x_4}}{a_1 \cdot a_2 \cdot a_3 \cdot a_4}$$

Diagrama de executie a programului este prezentata la adresa www.ivan.ase.ro/pseudo .

Programul care calculeaza amprenta textului utilizând numere pseudoaleatoare realizeaza urmatoarele:

- citeste textul;
- stabileste lungimea textului, vocabularului si alfabetului;
- numara cuvintele separate prin spatiu, punct, virgula;
- genereaza vectorul cu numere pseudoaleatoare;
- extrage esantionul de caractere;
- calculeaza frecventele de aparitie ale elementelor din alfabet;
- extrage esantionul de cuvinte;
- construieste vocabularul esantionului;
- calculeaza frecventele de aparitie a elementelor de vocabular;
- calculeaza numarul inversiunilor necesare sortarii frecventelor;
- normeaza numarul inversiunilor ;
- genereaza indicatorii a_1, a_2, a_3, a_4;
- calculeaza indicatorul agregat G;
- afiseaza valorile indicatorilor:
  - coeficientul de variatie al frecventelor caracterelor din alfabetul esantionului;
  - coeficientul de variatie al frecventelor cuvintelor din vocabularul esantionului;
  - numarul inversiunilor necesare sortarii frecventelor caracterelor din alfabetul esantionului, normat;
  - numarul inversiunilor necesare sortarii frecventelor cuvintelor din vocabularul esantionului, normat;
  - indicatorul agregat G.

Exemple privind utilizarea si testarea produsului software de stabilire a amprentei unui text prin generare de numere aleatoare sunt prezentate la adresa www.ivan.ase.ro/pseudo. Utilizarea produsului software impune executarea urmatorilor pasi:

1. se alege samânta de generare a numerelor pseudoaleatoare;
2. se încarca fisierul text de analizat;
3. se vizualizeaza alfabetul si vocabularul esantionului (optional).

Rularea în cascada a programului pe esantioanele succesiv obtinute va conduce la stabilizarea amprentei agregate într-un numar finit de pasi.

Concluzii

Fiecarui text i se calculeaza G si cei patru indicatori, construindu-se o baza de amprente.

Nume fisier text	**G**	**x_1**	**x_2**	**x_3**	**x_4**

Pentru analiza comparata a textelor se compara cei cinci indicatori. Daca indicatorii sunt egali, rezulta ca unul dintre texte este clona celuilalt si se procedeaza la analiza pe text direct. Software-ul trebuie sa asigure reproductibilitatea procesului de obtinere a amprentei. Acelasi text are tot timpul aceeasi amprenta în acelasi ipostaze de lucru.

Bibliografie

[IVAN02a] Ivan Ion, Niculescu Silviu, Catalin Boja – *Clonarea bazelor de date*, Revista Româna de Informatica si Automatica, Bucuresti, vol. 12, nr. 4, 2002, pg. 46 - 53

[IVAN02b] Ivan Ion, Pocatilu Paul, Popa Marius, Sacala Mihai, Ungureanu Doru – *Information Cloning*, Probleme regionale în contextul procesului de globalizare – Simpozion International, Chisinau, Republica Moldova, 9 – 10 octombrie 2002, pg. 371 – 375

[IVAN02c] Ivan Ion, Popa Marius, Sacala Mihai – *Ortogonalitatea datelor*, Revista Româna se Statistica, Bucuresti, nr. 4, 2002

[POCO03] Ivan Ion, Adrian Pocovnicu – *Construirea amprentei textului prin esantionare bazata pe generarea de numere pseudoaleatoare*, Bucuresti, 2003

[SMEU01] Smeureanu Ion, Dârdala Marian – *Programarea în limbajul C/C++*, Editura CISON, Bucuresti, 2001

[PORO93] Porojan Dumitru – *Statistica si teoria sondajului*, Casa de editura si presa "Sansa" SRL, Bucuresti, 1993

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