Text Entities Metrics

Asist. Marius POPA
Catedra de Informatică Economică, A.S.E. Bucureşti

This paper presents the characteristics of the models assigned to the metrics and the way in which the text entities metrics are developed. Also, metrics classes are identified and presented. It is built a text entity metric system for the following text entity quality characteristics: length, complexity and orthogonality.
Keywords: *metrics, text entity, quality characteristic.*

Modele asociate metricilor

O metrică este un model matematic de forma: $y = f(x_1, x_2, ..., x_{nft})$, unde:

- y este un model ce depinde de valorile x_1, x_2, …, x_{nft} ale factorilor Ft_1, Ft_2, …, Ft_{nft};
- x_i este valoarea numerică a factorului de influenţă Ft_i;
- Ft_i reprezintă factorul de influenţă i din mulţimea factorilor ce determină variabila rezultativă y.

Modelul matematic y este format din una sau mai multe ecuaţii sau inecuaţii şi conţine una sau mai multe funcţii obiectiv.

Se identifică existenţa următoarelor metrici, în funcţie de natura legăturilor dintre factorii de influenţă: metrici cu factori de influenţă direcţi şi metrici cu factori de influenţă indirecţi.

În [IVAN99a] sunt evidenţiate elemente care trebuie considerate la construirea unei metrici:

- identificarea factorilor de influenţă;
- măsurarea intensităţilor dintre fiecare factor şi variabila endogenă y;
- reprezentarea grafului dependenţelor;
- selectarea factorilor cu influenţă semnificativă;
- identificarea formelor analitice a dependenţelor;
- efectuarea de teste statistice în vederea selectării factorilor de influenţă.

Pentru evaluarea unei entităţi text, din mulţimea indicatorilor construiţi se selectează doar cei care sunt semnificativi în raport cu obiectivele analizei. Determinarea valorii indicatorilor asociaţi anumitor caracteristici pentru entităţile text analizate presupune parcurgerea unor etape:

- alegerea caracteristicilor de calitate în raport cu care entităţile sunt cuantificate numeric;
- cuantificarea caracteristicilor de calitate fundamentale ale entităţilor text;
- prelucrarea primară a datelor culese, observate;
- gruparea informaţiilor obţinute;
- agregarea datelor individuale.

Prin intermediul unui indicator, sunt evidenţiate expresii numerice ale obiectelor, fiinţelor, fenomenelor, proceselor, activităţi de orice natură, categorii economice şi sociale în funcţie de anumite criterii şi/sau caracteristici.

În funcţie de momentul obţinerii expresiilor numerice, indicatorii se clasifică în indicatori primari şi indicatori derivaţi.

Indicatorii primari sunt cei care se obţin în procesul de cuantificare numerică a entităţilor text pe baza caracteristicilor de calitate luate în considerare. În categoria indicatorilor primari intră:

- lungimea alfabetului utilizat în construcţia entităţii;
- lungimea mulţimii de separatori;
- lungimea vocabularului textului;
- frecvenţa de apariţie a simbolurilor din alfabet;
- frecvenţa de apariţie a cuvintelor din vocabularul textului;
- lungimea entităţii exprimată ca: număr de caractere, număr de cuvinte, numărul de pagini, numărul de octeţi ocupaţi de fişier etc.;
- numărul de operanzi şi operatori, în cazul entităţilor de tip programe sursă;

- numărul de concepte noi definite în cadrul domeniului abordat.

Numărul de indicatori primari este unul foarte mare, având în vedere multitudinea de criterii şi caracteristici de calitate care se iau în considerare în cadrul unei analize;

Metricile derivate sunt calculate pe baza indicatorilor primari, având asociate modele de evaluare a calităţii cu o complexitate ridicată.

Prin intermediul indicatorilor derivaţi sunt evidenţiate anumite aspecte calitative privind [ISAI95]:

- relaţiile dintre părţile unei colectivităţi;
- relaţiile dintre caracteristicile unei entităţi;
- relaţiile dintre fenomene;
- măsura influenţei anumitor factori asupra declanşării şi derulării unor procese şi fenomene.

În categoria mărimilor derivate sunt incluse: mărimile relative, mărimile medii, indicatorii variaţiei, indicii şi indicatorii ce caracterizează corelaţia. Categoria cea mai simplă a indicatorilor derivaţi cu cea mai mare răspândire este cea a mărimilor relative.

Metrici ale lungimii entităţilor text

Un text $T = <c_1\ c_2\ \ldots\ c_n>$ format din cuvintele c_1, c_2, …, c_n are lungimea L definită ca număr de cuvinte, Lgt(T) = n sau ca număr de caractere $Lgts(T) = \sum_{i=1}^{n} L(c_i)$, unde $L(c_i)$ este funcţia care dă numărul de caractere care alcătuiesc un cuvânt.

De exemplu, pentru textul:

$$T_1 = < aaa\ bbb\ ccc\ aaa >$$

se determină indicatori privind:

- lungimea textului exprimată ca număr de cuvinte este: $Lgt(T_1) = 4$ cuvinte;
- lungimea textului exprimată ca număr de caractere: $Lgts(T_1) = 12$ caractere;
- lungimea textului exprimată ca număr de separatori: $Lgtp(T_1) = 3$ separatori;
- lungimea textului exprimată ca număr de simboluri utilizate este:

$$LgtS(T_1) = Lgts(T_1) + Lgtp(T_1) = 15 \text{ simboluri}$$

- lungimea vocabularului exprimată ca număr de cuvinte este: $Lgv(V_A) = 3$ cuvinte, unde $V_A = <$ aaa, bbb, ccc $>$;
- lungimea textului exprimată ca spaţiul de memorie ocupat pe un suport electronic de stocare este:
  - cuvintele sunt codificate ASCII: $Lgtm(T_1) = 15$ octeţi;
  - cuvintele sunt codificate conform Unicode: $Lgtm(T_1) = 30$ octeţi.
- lungimea textului exprimată ca număr de construcţii sintactice aflate pe diferite niveluri de agregare; astfel, lungimea textelor se exprimă ca: număr de propoziţii, număr de paragrafe, număr de subcapitole, număr de capitole, număr de volume.

Transpunerea textelor în limbaj matematic formalizat are ca punct de plecare elementele componente ale acestora: alfabet, cuvânt, separatori, vocabular. Pornind de la aceste aspecte se definesc indicatori de măsurare a lungimii textului şi a vocabularului.

Pentru determinarea lungimii textului, există mai multe variante de exprimare a indicatorilor menţionaţi anterior. Astfel, modelele asociate indicatorilor sunt:

- lungimea textului exprimată ca număr de simboluri ale alfabetului care sunt utilizate în construirea cuvintelor se determină conform expresiei:

$$Lgts(T) = \sum_{i=1}^{ns} a_i \text{, unde:}$$

  - Lgts(T) este lungimea textului exprimată ca număr de simboluri din alfabetul A;
  - a_i reprezintă simbolul alfabetului de rang i dintr-un text;
  - ns reprezintă numărul de simboluri ale alfabetului A utilizate în construirea textului.
- lungimea textului exprimată ca număr de simboluri ale alfabetului la care se adaugă numărul de separatori utilizaţi în construirea propoziţiilor şi frazelor:

$$LgtS(T) = \sum_{i=1}^{ns} a_i + \sum_{j=1}^{r} s_j \text{, unde:}$$

  - LgtS(T) este lungimea textului exprimată ca număr de simboluri;
  - a_i reprezintă caracterul de rang i dintr-un text;
  - s_j este separatorul de rang j dintr-un text;
  - ns reprezintă numărul de simboluri ale alfabetului utilizate în construirea textului;
  - r este numărul de separatori folosiţi în

text.

- lungimea textului exprimată ca număr de cuvinte: $Lgt(T)=\sum_{i=1}^{n} c_i$, unde:

  - Lgt(T) este lungimea textului exprimată în număr de cuvinte;
  - c_i reprezintă cuvântul cu rangul i din textul T;
  - n este numărul de cuvinte din vocabularul V_A folosite în text.

Având în vedere faptul că un cuvânt este format din simboluri ce alcătuiesc alfabetul A, lungimea unui text exprimată ca număr de caractere se determină conform expresiei:

$$Lgts(T)=\sum_{i=1}^{n} L(c_i)$$

- frecvența de apariție a unui cuvânt într-un text reprezintă numărul de apariții; astfel, lungimea textului T exprimată ca număr de cuvinte se determină pe baza formulei:

$$Lgt(T)=\sum_{i=1}^{ncv} f_i \text{ , unde:}$$

  - Lgt(T) este lungimea textului exprimată în număr de cuvinte;
  - f_i reprezintă frecvența de apariție în text a cuvântului i din vocabularul V_A;
  - ncv este lungimea vocabularului V_A.

De asemenea, pe baza frecvențelor de apariție se determină lungimea textului exprimată ca număr de caractere conform relației:

$$Lgts(T)=\sum_{i=1}^{ns} fs_i \text{ , unde:}$$

  - Lgts(T) este lungimea textului exprimată în număr de caractere din alfabetul A;
  - fs_i reprezintă frecvența de apariție în text a simbolului i din alfabetul A;
  - ns este numărul de simboluri din alfabetul A.

- lungimea medie a unui cuvânt din textul T se determină conform expresiei:

$$LgMc=\frac{Lgts(T)}{Lgt(T)} \text{, unde:}$$

  - LgMc este lungimea medie a unui cuvânt din textul T;
  - Lgts(T) este lungimea textului exprimată ca număr de caractere;
  - Lgt(T) reprezintă lungimea textului T exprimată ca număr de cuvinte.

Rezultă că lungimea textului T exprimată ca număr de caractere este:

$$Lgts(T) = Lgt(T) \cdot LgMc$$

Există mai multe niveluri de agregare a construcțiilor sintactice în funcție de care sunt dezvoltate modele asociate metricilor ce privesc lungimea entităților text. Astfel, pentru o metrică se asociază mai multe modele.

Metrici ale complexității entităților text

Lumea reală este formată din obiecte, ființe, fenomene şi procese care prezintă caracteristici de spațiu, timp, natură etc. Descrierea lumii reale se realizează prin intermediul entităților text. Entitatea text este un sistem şi transpune într-o formă scrisă complexitatea sistemelor din lumea reală.

Complexitatea priveşte nu numai natura elementelor şi relațiile acestora dintr-un sistem, ci şi volumul şi dificultatea cuantificărilor numerice a anumitor caracteristici ale entităților, respectiv a sistemelor în general.

Complexitatea entităților text vizează împărțirea acestora în construcții de ordin sintactic şi relațiile dintre părțile componente. Astfel, complexitatea entităților text este abordată pe două niveluri:

- numărul nivelurilor de descompunere a entităților;
- volumul structurilor componente pe un anumit nivel.

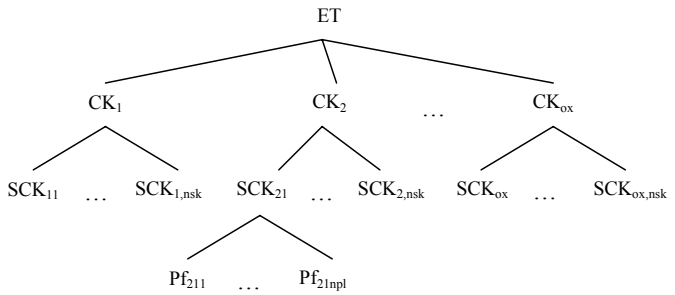

Fig. 1. Structura arborescentă a unei entități text

unde:

- ET este entitate text;
- CK_i reprezintă capitolul i din ET;
- SCK_{ij} este subcapitolul j din capitolul CK_i;
- Pf_{ijk} reprezintă paragraful k din SCK_j ce aparţine lui CK_i;

Complexitatea datelor rezultă din agregarea numărului de apariţii şi numărului de legături dintre ele, după relaţia [IVAN99b]:

$$Cx(DT) = noz \log_2 noz + nop \log_2 nop$$

unde:

- noz este numărul de apariţii;
- nop reprezintă numărul de operatori.

Determinarea complexităţii unei entităţi text presupune luarea în considerare a următoarelor elemente:

- numărul cuvintelor de bază care alcătuiesc textul, noz;
- relaţiile dintre cuvinte exprimate prin numărul de cuvinte operatori, nop; sunt exprimate prin intermediul verbelor;
- numărul cuvintelor de legătură, nlg; se exprimă prin prepoziţii şi conjuncţii.

Expresia analitică a modelului pe baza căruia se determină măsura complexităţii este:

$$Cx(T) = noz \log_2 noz + nop \log_2 nop + n\lg \log_2 n\lg$$

Se consideră textul:

T = < Metrica informaţiei joacă un rol foarte important pentru măsurarea diversităţii >

Pentru textul T, valorile operanzilor indicatorului Cx(T) sunt:

- noz = 6 cuvinte, corespunzător cuvintelor: *metrica, informaţie, rol, important, măsurare, diversitate*;
- nop = 1 cuvânt, corespunzător verbului *a juca*;
- nlg = 3 cuvinte, corespunzător cuvintelor *un, foarte, pentru*.

Valoarea indicatorului Cx(T) = 20,25. Complexitatea Cx(T) vizează construcţiile sintactice componente ale unei entităţi text.

Măsurarea complexităţii relaţiilor dintre construcţiile de ordin sintactic ale entităţii text ET are loc pe baza matricei de precedenţe MP prin intermediul căreia sunt evidenţiate relaţiile de dependenţă, ordonare, aranjare a construcţiilor sintactice. Pentru determinarea complexităţii se iau în considerare elementele situate în matricea precedenţelor. Complexitatea are următoarea expresie analitică:

$$Cx(MP) = \sum_{i=1}^{ncv} ((\sum_{j=1}^{ncv} mc_{ij}) \log_2 (\sum_{j=1}^{ncv} mc_{ij})), \text{ unde:}$$

- ncv este numărul de cuvinte din vocabularul textului;
- mc_{ij} reprezintă elementul de pe linia i, coloana j din matricea precedenţelor.

Determinarea complexităţii sistemului de caracteristici permite efectuarea de analize comparative cu alte entităţi text şi poziţionarea sa în raport cu acestea. Evaluarea complexităţii se bazează pe cuantificarea numerică a construcţiilor de ordin sintactic şi a legăturilor dintre acestea.

Metrici ale ortogonalităţii textelor

Ortogonalitatea evidenţiază diferenţele dintre două entităţi, indiferent de forma de reprezentarea a acestora. Datele sunt ortogonale dacă ele sunt complet diferite, [IVAN04]. Se defineşte un indicator de ortogonalitate care, în caz de ortogonalitate ia valoarea 1, iar în caz de identitate ia valoarea 0. Astfel, se remarcă existenţa unui spectru al gradelor de ortogonalitate dintre două elemente, cu valori în intervalul [0,1].

Se consideră textele T_1, T_2, ..., T_{nx}, din care se extrag textele T_i şi T_j. Se dezvoltă şi se implementează modele asociate metricilor de ortogonalitate în vederea comparării entităţilor text.

În [IVAN02], se optimizează indicatorul de ortogonalitate a două texte, O_{Lij}, având relevanţă numai valorile pozitive subunitare ale acestuia:

$$O_{L_{ij}} = \frac{\min(L_i, L_j)}{\max(L_i, L_j)}$$

De asemenea, sunt definiţi indicatori de asemănare a textelor pornind de la elementele structurale ale acestora. Se consideră două texte T_i şi T_j. Aplicarea operaţiei de reuniune a vocabularelor celor două texte conduce la obţinerea vocabularului:

$$V_{re} = VT_i \; U \; VT_j$$

Pentru vocabularul reuniune V_{re}, ncvr repre-

zintă numărul de cuvinte. Comparația a două texte, luându-se în considerare frecvențele de apariție a cuvintelor utilizate, este realizată prin intermediul indicatorului următor, [IVAN02]: $O_V = \frac{\sum_{i=1}^{ncvr} nct_i}{ncvr}$, unde nct_i reprezintă indicatorul care ia una din următoarele valori:

- 0, dacă frecvențele de apariție pentru cuvântul c_i sunt diferite sau cel puțin una din ele este nulă;
- 1, dacă pentru cuvântul c_i frecvențele de apariție sunt identice în cele două texte.

Măsura în care două texte utilizează același fond de cuvinte al unui vocabular este dată prin calcularea următorului indicator, [IVAN03a]:

$$O_{FC} = \frac{\min\{Lgv(V_{co}), Lgv(V_{re})\}}{\max\{Lgv(V_{co}), Lgv(V_{re})\}}$$

unde:

- V_{co} – vocabularului de cuvinte comune celor două texte;
- V_{re} – vocabularului reuniune al textelor supuse studiului comparativ.

Ponderea elementelor identice în două texte este dată de raportul, [IVAN03a]:

$$O_{PI} = \frac{ncrf}{Lgv(V_{re})}$$

unde ncrf este numărul de cuvinte din vocabularul reuniune care au aceeași frecvență de apariție.

În cazul entităților text de tip program sursă, ET_S, în [IVAN04] sunt definite metrici asociate caracteristicilor unor programe de acest tip. Sunt definite metrici de ortogonalitate pe baza acestor caracteristici. Pentru două texte sursă ET_{Si} și ET_{Sj}, forma generală a indicatorului de ortogonalitate pentru metrica m este:

$$O_{ij} = \frac{\min(M_i^m, M_j^m)}{\max(M_i^m, M_j^m)}, \text{ unde:}$$

- M_i^m reprezintă valoarea metricii m pentru programul sursă ET_{Si};
- M_j^m este valoarea metricii m pentru programul sursă ET_{Sj}.

Definirea metricilor pe texte oferă o imagine numerică a acestora și permite efectuarea de analize cantitative.

Concluzii

Datele dețin un rol foarte important în reprezentarea realității. Datele sunt culese, stocate, prelucrate, fundamentează decizii și determină acțiuni. Competitivitatea unei activități este dată de unul din cei mai importanți indicatori, respectiv calitatea.

Stabilirea nivelului de calitate a datelor reprezentate sub formă de text se realizează prin intermediul sistemului de metrici construit. Pe baza valorilor calculate se fundamentează decizii privind asigurarea și creșterea calității textelor. Din mulțimea indicatorilor construiți, se detașează ca importanță cei care măsoară ortogonalitatea. Astfel, sunt evidențiate contribuțiile originale la dezvoltarea domeniului în care entitatea text este dezvoltată.

Bibliografie

[IVAN05] Ion IVAN, Marius POPA – *Uniformity – Quality Characteristic of Text Entities,* în „Informatica Economică", vol. 9, nr. 1, 2005, pg. 84 – 88

[IVAN04] Ion IVAN, Marius POPA – *Ortogonalitatea produselor program orientate obiect*, în „Informatica Economică", vol. 8, nr. 4, 2004, pg. 93 – 96

[IVAN03a] Ion IVAN, Marius POPA, Paul POCATILU – *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

[IVAN03b] Ion IVAN, Marius POPA, Sergiu CAPISIZU, Lukacs BREDA, Bogdan FLORESCU – *Clonarea informatică*, Editura ASE, Bucureşti, 2003

[IVAN02] Ion IVAN, Marius POPA, Mihai SACALĂ – *Ortogonalitatea datelor*, în "Revista Română de Statistică", vol. 11, nr. 4, 2002, pg. 30 - 45

[IVAN99a] Ion IVAN, Panagiotis SINIOROS, Mihai POPESCU, Felix SIMION – *Metrici software*, Editura INFOREC, Bucureşti, 1999

[IVAN99b] Ion IVAN, Gheorghe NOŞCA, Sebastian TCACIUC, Otilia PÂRLOG, Răzvan CĂCIULĂ – *Calitatea datelor*, Editura INFOREC, Bucureşti, 1999

[ISAI95] Alexandru Isaic-Maniu, Constantin Mitruț, Vergil Voineagu – Statistica pentru managementul afacerilor, Editura Economică, Bucureşti, 1999

[WANG95] Richard WANG, M. REDDY, Henry KON – *Toward Quality Data: An Attribute-based approach*, în „Decision Support Systems", nr. 13, 1995, pg. 349 – 372

[www1] www.hci.com.au

[www2] www.research.ibm.com