EVALUATION METHODS OF THE TEXT ENTITIES

Marius POPA[1]
PhD, University Lecturer, Economic Informatics Department
Academy of Economic Studies, Bucharest, Romania
PhD Disertation Title:
Methods and techniques used in the qualitative evaluation of the data (2005)

E-mail: marius.popa@ase.ro

Abstract: *The paper highlights some evaluation methods to assess the quality characteristics of the text entities. The main concepts used in building and evaluation processes of the text entities are presented. Also, some aggregated metrics for orthogonality measurements are presented. The evaluation process for automatic evaluation of the text entities is made by software application. These ones implement the metric system for text entity quality characteristic evaluation. The metrics and software application are validated through testing examples.*

Key words: Assessment, Quality, Text entity

1. Introduction: Concepts and definitions

In (Marius Popa, 2005), (Ivan, Popa, 2005), (Ivan, 2003), (Ivan, Popa, Boja, Toma, 2005, 43–57) some concepts used in building, analysis and evaluation of the text entities are defined and presented. The used concepts include the following elements: alphabet, word, vocabulary, subvocabulary, text, template, structured text, entity.

For each used concept, it is offered a definition, the necessity of its using, its characteristics, using forms, ways for information representation through its using, models, requirements and examples of building and using.

Through defined concepts, it is highlighted some representation and structuring forms of data. The data quality is given by the level assured for quality characteristics associated to data. The identification and quantification of data quality characteristics are critical activities in control and assurance processes of the quality.

According to definition from (Marius Popa, 2005), the text entities are constructions formed by word strings characterized by word positions in text, word grouping in order to define a context, by correspondence of the words with elements, actions and phenomenon from real world, qualitative attributes that group concrete aspects from real world in homogenous collectivities in connection with established criteria.

In (Ivan, Popa, 2005), the conditions that must be respected in building process of the text entities are established. These ones look upon the following aspects:

- A strong delimitation of the approached domain;
- Key word definition for the domain;
- Vocabulary used, that also includes the key word vocabulary;

- Concepts, techniques, methods, methodology and technology knowledge for the domain;
- Detail and other domain connected element documentation;
- Respecting of syntax rules for each language;
- Rules to be followed, regarding entity structuring, progressive approaching of the problems, usage of standard formats to represent the text information.

In (Department of Defence 8320.1-M, 1994), in accordance with Federal Information Processing Standards from United States of America, the data quality is defined as accuracy, opportunity, completeness, importance and accessibility that make data to be appropriate, that is to be corresponding with its usage.

Data quality includes the activity and data model usage, entities, attributes, metadata, diagrams and data architectures. The text entity quality ET is given by all features that the entity have. These ones are perceived and appreciated by the persons whom are part of a group. In comparison with an evaluation system, appropriate for each person, the text entity ET has associated a score, a mark that differentiates it of other text entities or includes it in a collection (Ivan, Popa, 2005).

2. Evaluation formulae

Aspects regarding the model development associated with evaluation metrics of the text entity quality characteristics are presented in (Popa, 2005), (Ivan, Popa, 2005), (Ivan, 2003), (Ivan, Popa, Boja, Toma, 2005, 43–57.

The evaluation metric building of text entities leads to text quality evaluation system making. The metrics included in this system are structured in two classes function of complexity classes of the used concepts as result of their aggregation:

- Quality characteristic metrics developed on the base of structure and semantic content of text entities;
- Metrics of the data representation form on the base on a representation reference system.

In first category of metrics, in (Marius Popa, 2005) metrics regarding the volume and dynamics of data, correctness, completeness, reliability, complexity, comparability, homogeneity and orthogonality of text entities were developed. It remarks as importance the sub-category of metrics developed in order to measure the orthogonality.

In the second metric class, there are included the quantification models for fundamental syntactical construction orthogonality used for text entity building. These constructions aim: symbol, character and word.

In (Marius Popa, 2005), a reference system is defined in order to represent the symbols from the alphabet. Metrics for the alphabet internal orthogonality evaluation are developed and also metrics associated to the orthogonality among alphabets.

In order to determine the orthogonality between two symbols a_i and a_j of a alphabet, it is built the metric $H(a_i, a_j)$. A main importance to make conclusions regarding the whole alphabet symbol orthogonality is given by aggregated indicator computation with the following analytical form (Marius Popa, 2005):

$$\overline{H}(A_L)=\sqrt[\frac{n(n-1)}{2}]{\prod_{i=1}^{n-1}(\prod_{j=i+1}^{n}H(a_i,a_j))}$$

where n represents the symbol number of the alphabet.

In the same category of metrics, there are include the metrics that measure the word orthogonality. Thus, there are determined the words that are part of the same word family, identifying the words with the same root. In (Marius Popa, 2005), methods and models form word family identification are presented. Also, aggregation processes of the primary indicator values are implemented.

In the most part of the cases, the indicator aggregation is made by geometrical mean using. This thing is favoured by the fact that the orthogonality indicator values can be structured on two dimensions, what leads to a metric with the following analytical form:

$$\gamma_f=\sqrt[C_n^2]{\prod_{i=1}^{C_n^2}\gamma_{pi}}$$

where:

γ_f – aggregated metric for orthogonality evaluation;

C_n^2 - value number resulted from orthogonality metric applying among different text constructions;

γ_{pi} – primary metric for orthogonality evaluation.

On the base of aggregated metrics, conclusions regarding the characteristic for the whole collectivity are obtained.

3. Evaluation algorithms

Using of a text entity evaluation metric doesn't suppose anytime only the proper model using, but requirements assurance for the input data.

The bigger complexity of the models associated to text entity evaluation metrics determines the algorithm development for input data preparing, model implementation and post-evaluation operations to permit a big accuracy interpretation of the characteristics measured by metric.

Thus, in (Marius Popa, 2005), (Ion Ivan, Daniel Milodin, Marius Popa, 2005, 41–56) there are developed and implemented algorithms for implementation of the models associated to text entity metric. For example, for metric quantification $H(a_i, a_j)$ regarding two symbol orthogonality from an alphabet the following algorithm was developed and implemented:

P1: it is defined a reference system formed by the segments s_1, s_2, ..., s_{ns} used to build each symbol from the alphabet;; ns represents the segment number from considered reference system.

P2: it defined a reference rule of the reference system segments.

P3: it associates a rank r_i for each segment s_i from reference system, obtaining the pairs (s_i, r_i).

P4: it represents the alphabet symbols, using the reference system.

P5: it builds a matrix $M(A_L)$ such as the element m_{ij} = 1 if to build the symbol a_i from the alphabet A_L it uses the reference segment s_j. If the reference segment s_j is not used then m_{ij} = 0.
P6: it computes the sums on columns, S_j, to obtain the using frequencies of the segments from the reference system in symbol defining from the alphabet.
P7: it computes the maximum and minimum sums, S_{max} and S_{min}.
P8: it normalizes the values S_j on the base of the expression:

$$Sn_j = \frac{S_{\max} - S_j}{S_{\max} - S_{\min}}$$

The values Sn_j are included in [0; 1].
P9: it interchanges the columns of the matrix $M(A_L)$ to obtain an ascendant order for the values Sn_j.
P10: it makes the correspondence of the values Sn_j with the segments s_j from the chosen reference system.
P11: it re-codifying the ranks of the reference system such as the new numbers to highlight the using frequencies, obtaining the pairs *(s_i, $r_i^{'}$)*.

The presented algorithm is a rigorous way to evaluate the orthogonality of the symbol representations in an alphabet. The symbol representation orthogonality increasing has importance and use in building process of the text entity with symbols good differentiated.

4. Evaluation software

The determination through a software application of the quality characteristic values and text entity orthogonality metrics suppose the carrying on of the following activities:

- Application objective definition;
- Input establishment on the base of quality characteristic system and metric model study.
- System architecture building;
- Collecting, normalizing and organizing of the data in correspondence with metric requirements;
- Metric system implementation;
- User interface designing in assistance of the process to establish the text entity base orthogonality;
- Metric system testing, tracing the software product behavior in limit cases especially.

In (Marius Popa, 2002) is presented the architecture and function of the product *Cloning Analysis Software – CAS*. This software application implements the metrics for the fundamental characteristics for texts and data organized in matrixes. In figure 1, there are highlighted the modules of CAS application.

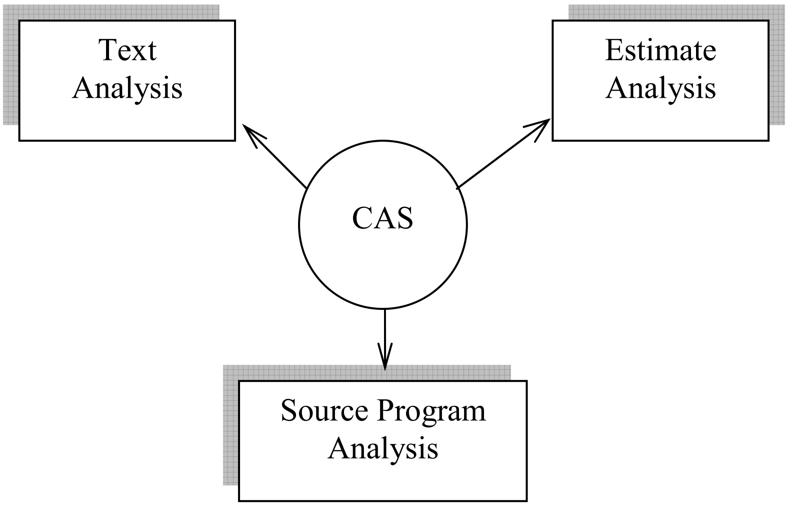

Figure 1. The modules of CAS application

The text orthogonality analysis from the text entity base supposes the building of aggregated orthogonality indicators matrix for the text entity pairs.

The orthogonality aggregated indicator associated to entity pair is obtained through orthogonality determination for the following primary metrics, (Marius Popa, 2002): entity length, appearance frequencies of the alphabetic characters, user vocabulary, text entity vocabulary, common vocabulary, the entity structure.

In the quantitative analysis of the estimates, the used algorithms for orthogonality aggregated indicator suppose the following step passing:

a. Data structure initialization loaded with data about estimates;
b. Data loaded regarding the estimate structure and their content;
c. Derived value determination;
d. Comparison of primary and derived values;
e. Orthogonality aggregated indicator determination.

The source program orthogonality analysis from the project annexes of the text entity bases supposes the building of the aggregated orthogonality indicator matrix. The aggregated indicator is obtained through the following primary metric determination, (Marius Popa, 2002): program length, appearance frequencies of the alphabetical characters, user vocabulary, program vocabulary, common vocabulary, entity structure, defined variables, precedent matrix of the variables, variable position.

In (Marius Popa, 2005), there are presents the application characteristics of *Bibliography Analysis – BA* that performs regarding: appearance frequencies of the syntactical constructions, file structuring, measuring of the word finding degree, key word searching, bibliography elements processing. The application offers to the user some aggregated indicators regarding the analyzed elements.

Software automatizes the evaluation process of the text entity quality, a very important aspect in their qualitative analysis.

5. Testing examples

It considers the Slav alphabet A_S and Greek alphabet A_G. The internal orthogonality indicator evaluation algorithm depending of representation way, there are obtained values structured in a matrix structure.

Thus, the appearance matrix of the values for symbol pair orthogonality of the Slav alphabet is presented in table 1.

Table 1. The frequencies of the orthogonality for the alphabet A_S

Value	Frequency	Value	Frequency	Value	Frequency	Value	Frequency
1,00	154	0,63	22	0,20	8	0,13	2
0,50	74	0,78	22	0,86	8	0,22	2
0,80	72	0,25	20	0,88	8	0,10	2
0,60	66	0,56	20	0,44	8	0,85	2
0,75	66	0,82	20	0,77	6	0,11	2
0,67	46	0,91	18	0,55	6	0,58	2
0,83	40	0,90	18	0,64	6	0,79	2
0,00	35	0,92	18	0,45	4	0,36	2
0,40	32	0,89	14	0,29	4	0,43	2
0,70	30	0,38	12	0,14	4		
0,73	30	0,71	12	0,17	4		
0,33	22	0,30	10	0,57	4		

The values of the orthogonality levels for symbol pairs from the Greek alphabet are highlighted in the following table:

Table 2. The frequencies of the orthogonality for the alphabet A_G

Value	Frequency	Value	Frequency	Value	Frequency	Value	Frequency
1,00	200	0,86	14	0,90	8	0,38	4
0,75	56	0,88	14	0,73	6	0,10	2
0,80	42	0,78	10	0,85	6	0,11	2
0,50	36	0,40	10	0,20	6	0,42	2
0,67	34	0,71	10	0,29	6	0,43	2
0,60	30	0,91	10	0,63	6	0,25	2
0,00	27	0,64	8	0,89	4	0,58	2
0,83	24	0,33	8	0,56	4	0,92	2
0,82	16	0,70	8	0,30	4		

The symbol representation is made on the base of the reference system from (Ion Ivan, Daniel Milodin, Marius Popa, 2005, 41–56). In (Ion Ivan, Daniel Milodin, Marius Popa, 2005, 41–56), a comparative analysis of the alphabet orthogonality is made on the base of the values included in matrixes with the orthogonality values appearance frequencies.

The aggregated values of the internal orthogonality of the two alphabets are given in table 3.

Table 3. Alphabet orthogonality indicator values

	Slav alphabet	Greek alphabet
Internal Orthogonality	0,66	0,75

The orthogonality analysis permits the alphabet design that increases the orthogonality. The character representation orthogonality increasing is important because the symbols from the alphabet have a better differentiation.

Conclusions

The paper highlights some techniques and methods for text entity evaluation. The emphasis is on orthogonality characteristic that allows the qualitative improvements in building and evaluation processes for the text entities.

The software products have a plus of efficiency in order to get to proposed objectives, and the testing examples contribute to proposed algorithm and developed software application validation

Bibliography

1. Anany Levitin, Thomas Redman, **Data as a Resource: Properties, Implications, and Prescriptions**, Sloan Management Review, 1998, p. 89–101
2. Department of Defense 8320.1-M, **Data Quality Assurance Procedures (Draft),** Quality Information for a Strong Defense, Department of Defense, 1994
3. Ion Ivan, Marius Popa, Text Entities **Development, Evaluation, Analysis,** Bucharest: ASE Printing House, 2005
4. Ion Ivan et al., **Information Cloning,** Bucharest: ASE Printing House, 2003
5. Ion Ivan, Marius Popa, Cătălin Boja, Cristian Toma **Metrics for Text Entity Similarity**, Studii şi Cercetări de Calcul Economic şi Cibernetică Economică 39, no. 4, 2005, p. 43–57
6. Ion Ivan, Daniel Milodin, Marius Popa, **Alphabet Orthogonality**, Revista Română de Informatică şi Automatică 15, no. 3, 2005, p. 41–56
7. Ion Ivan, Marius Popa, Alexandru Popescu **The Aggregation of the Text Entities**, Economic Computation and Economic Cybernetics Studies and Research 38, no. 1-4, 2004, p. 37–50
8. Ion Ivan, Marius Popa **Text Metric Type,** Studii şi Cercetări de Calcul Economic şi Cibernetică Economică 38, no. 1, 2004, p. 25–36
9. Ion Ivan, Marius Popa, Constantin Avram **Similarity Degree Using in Software Fingerprint Building,** Studii şi Cercetări de Calcul Economic şi Cibernetică Economică 37, no. 2, 2003, p. 39–54
10. Leo Pipino, Yang W. Lee, Richard Y. Wang **Data Quality Assessment**, Comunications of the ACM 45, no. 4, 2002, p. 211–218
11. Marius Popa **Techniques and Methods for Data Quality Evaluation**, Ph.D. dissertation, ASE Bucharest, 2005
12. Marius Popa **Text Entity Quality Evaluation – Theory and Practice**, Bucharest: ASE Printing House, 2005
13. Marius Popa **Text Entities Metrics**, Informatica Economică 9, no. 2, 2005, p. 56–60
14. Marius Popa, **Software for Similarity Degree Measuring of the Archived Files**, Bachelor's Degree, ASE Bucharest, 2002

[1] Marius Popa is lecturer in Economic Informatics Department, Academy of Economic Studies of Bucharest. He published over 30 articles in journals and magazines in computer science, informatics and statistics fields, over 35 papers presented at national and international conferences, symposiums and workshops. He is the author of one book and he is coauthor of three books. In November 2005, he finished the doctoral stage, and his PhD thesis has the title Methods and techniques used in the qualitative evaluation of the data. His interest domains are: data and software quality, software engineering and project management.