Vocabulary Richness Metric for Extracting Author's Semantic Mark in English Written Literary Works

Mădălina ZURINI[1], Alin ZAMFIROIU[1,2]
[1]Bucharest University of Economic Studies, Romania
[2]The National Institute for Research & Development in Informatics
madalina.zurini@csie.ase.ro, zamfiroiu@ici.ro

The present paper starts from a short introduction of the major aspects debated regarding the stylometric measures used for extracting the personal signature added by a particular author to its English written works. Those measures are used in the context of indicating an author from a limited cardinality set of authors being given a set of documents or a defined indicators values which characterizes the semantic way that an author is writing its works. The paper addresses the problems of the semantic level of a work depending on the tokens that he uses in the paper, tokens that are extracted in a preprocessing step of analysis. The tokens are defined using a lexical ontology, for the English words referring to WordNet, and the automatic extracting of those tokens from the words found in the particular processed papers. The main vocabulary richness evaluation metrics are presented taking into account the major literature review and extracting the main steps into a new proposed metric that is combining the vocabulary richness with the semantic layer of a paper. The concept of author mark is described. The objective of this research paper is highlighted into the new proposed metric that is non-dependent on the main subject discussed in the analyzed paper. This objective leads to a general metric that combines documents from different subjects into a metric that can describe the vocabulary richness of a specific author depending on the works that he had written. Furthermore, the analysis is conducting into a time evolution of this metric, using the extraction of the trend of the author's vocabulary richness indicator. Using a set of 13 years values of this indicator upon a specific author, the results are presented in this research paper. Future work refers to inserting this metric into a general description of the author mark into his specific English written works.

Keywords: *Stylometry analysis, Metrics, Author Mark, Lexical Ontology, Time-Trend Analysis, Intrinsic Plagiarism Detection*

1 Introduction

Intrinsic plagiarism detection implies the recognition of those parts of text within a document that are different taking into account the writing style of a certain author. Those parts are later on analysed as input data for the verification using external plagiarism detection tools. If a document in written by a single author, it is supposed that the passages written by him to be similar accordingly to its unique writing style.

Using this technique of comparing the writing style within each part of text from the papers written by multiple authors and adding unsupervised automatic classification techniques, those parts of text are grouped in clusters depending on the membership of each author. The problem of plagiarism detection using this type of analysis involves extracting the unique writing style of each author, method also called stylometry analysis. Having a set of characteristics that best describes in a unique manner the writing style of an author, a metric is created for value description of percentage membership of documents to authors. In the research conducted in [3], [4], [5], [6], [7], [8] and [10], the problems and methods of inserting intrinsic plagiarism are referred, adding into the discussion also the stylometry, the writing style of a specific author over his history of research or just within a single document.

Regardless of the type of plagiarism evaluation, intrinsic or external one, it is very important to determine the set of characteristics

DOI: 10.12948/issn14531305/20.3.2016.04

that must be taken into account in order to obtain as accurate results as possible. Those characteristics depend on the set of analysed documents, the language in which the documents are written and also the type of documents. The present research paper addresses the problem of literary English written documents by English native or European authors. For extracting from the initial set of documents the semantic analysis that describes the stylometry, multidimensional analysis is used. Chapter 2 reveals the relation between semantic analysis and main vocabulary richness metrics used in order to extract a value indicator of the words found in the analysed authors' set of written documents, transformed into tokens, and the semantic distances between them. The terms of words, tokens and frequency appearance are presented along with the main set of features of written style. The pre-processing phase is also presented, a step needed to convert words into WordNet tokens.

In chapter 3, the improved semantic richness vocabulary metric is presented and defined along with an example of applying in upon a given phrase. The time evolution analyses is done within chapter 4, where 13 years values are inserted into a time series. Using three methods, absolute mean change, average index and linear regression, the trend indicator is evaluated. Comparing the sum of squared errors of the three methods, the linear regression method is chosen for the forecast. The conclusions are withdrawn in chapter 5 along with the future work directions.

2 Vocabulary Richness Metrics In Stilometry Analysis

For analysis of an author's style of writing in the context of external analysis or intrinsic characteristics of plagiarism, the richness of vocabulary is defined as the characteristic of the author defines the degree to which the author uses words in a wider or narrower vocabulary. This feature was demonstrated in works such as [1], [2], as a feature closely related to the author, it can be fed into optimal set of features of the style of writing.

Table 1 contains a list of metrics used to assess vocabulary wealth within the set of features writing, detailing the variables in formulas defined metrics that are presented in this paper [1].

Table 1. Formulas for assessing the richness of vocabulary within the set of features of writing style

Vocabulary richness metrics
$\boldsymbol{Type - Token = V/N}$
$\boldsymbol{K = 10^4 \left(\sum i^2 V_i - N \right) / N^2}$
$\boldsymbol{R = V / \sqrt{N}}$
$\boldsymbol{C = \log V / \log N}$
$\boldsymbol{H = (100 \log N) / (1 - V_1 / V)}$
$\boldsymbol{Entropy = -100 \sum p_v \log p_v}$

where:

N – total number of words in the document analyzed;

V – total concepts identified in the set of words;

V_i – total concepts that appear of i times in the document;

p_v – the relative frequency of the most v present concept in the document.

Preprocessing phase for drawing text vocabulary wealth consists in separating the words of the text or analyzed fragment text, eliminating spaces and punctuation. An optimization of the processing and disposal is given connecting words, which are present in any text written by different authors. Denoting with V the set of words resulting from the preprocessing phase, $W = \{w_1, w_2, \dots, w_i, \dots, w_N\}$, insert the analysis and ontology WordNet lexical set of unique concepts for generating recovered from the initial set of words W intersection concepts in WordNet by reducing duplication and creating vector occurrences of each concept found so:

$$\begin{cases} T = \{t_1, t_2, \dots, t_i, \dots, t_{nt}\} \\ nap = \{nap_1, nap_2, \dots, nap_i, \dots, nap_{nt}\} \end{cases}$$

where:

DOI: 10.12948/issn14531305/20.3.2016.04

- *T* represents the set of unique concepts identified in the text and found in ontology WordNet;
- *nap represent* the set made up of the number of occurrences of each concept from the *T* set in the analyzed document.

While most indicators measuring the wealth of vocabulary used by the author of the work refers to the relationship between the number of unique words identified in a analyzed text in relation to the total number of existing words in that text, these metrics do not account instead the existing semantic component derived from those specific words extracted. Also proposed metrics extracted from the literature and does not assess the time course of this feature is implemented in a very high percentage in assessing a person's writing style.

Starting from this issue, it needs metrics to evaluate the proposal while richness of vocabulary used in this document under review and in previous documents, if they exist. Metric uses the number of words found, WordNet lexical concepts identified using ontology extraction through processing of root words and functions for calculating distances between any two concepts from WordNet.

3 Improved Metric for Evaluating the Vocabulary Richness In The Presence Of Semantic Relations

Impact of using this metric is given by the semantic side added in the set of words used in an analyzed text. By enriching this metric with semantic analysis feature generates a complex stylometry, the local point of view and in terms of the time course.

Thus, *ISRV*, Indicator of Semantics Richness of Vocabulary, it is defined as being equal to:

$$ISRV = \frac{\sum_{i=1}^{nt} nap_i * d\max(t_i)}{N}$$

where:

- nap_i represent the number of unique instances of the word founded on the position i of the unique set of terms extracted from the analyzed document;
- nt represent the cardinality of the set of unique terms extracted from the analyzed document;
- $dmax(t_i)$ is the maximum distance between single term and any other single term extracted from the set of terms, $dmax(t_i) = \max_{\substack{k \\ k \neq i}} d(t_i, t_k)$, distance is calculated using semantic distances defined in the WordNet lexical ontology, [9];
- N is the cardinality of the set of words, single or not, extracted from the analyzed document resulting from the preprocessing phase of the text.

The indicator $ISRV \in [0;1]$, and a value of the distances $dmax(t_i) \to 0, \forall i = \overline{1,nt}$ lead to the indicator value $ISRV \to 0$. Interpretation of this context consists of a document which is composed of words, possible distinctive or not, but who find themselves in the same semantic area in terms of the distance of the semantic ontology WordNet.

The opposite situation, where the $dmax(t_i) \to 1, \forall i = \overline{1,nt}$, transforms proposed indicator signs consistent with existing literature and specialty used to measure wealth vocabulary used by an author in a text or fragment text.

Table 2 contains examples of running the proposed metrics to assess the initial results. To assess the distance between any two concepts within the ontology Word-Net is used metric type Path Length, $d_{PATH}(c_1, c_2) = \frac{1}{lg(c_1, c_2)}$, metric that takes values in the range $[0; 1]$. It is also made a comparative analysis of the proposed metric type Type - Token metric format generally accepted assessment wealth vocabulary.

The metric proposed, based on metrics type Unique concept - Words and weighted distances semantic statements based on ontology WordNet added, besides the transformation of words into concepts WordNet, and distances semantic maximum of words, generating a component semantic not inserted in research on evaluation, measuring and interpreting the wealth of vocabulary used by an author in a text or piece of text written in English.

DOI: 10.12948/issn14531305/20.3.2016.04

Table Error! No text of specified style in document.**.** ISRV metric rolling on a set of testing compared to standard metric Type - Token

Analyzed fragment text	Vocabulary richness metrics are in depth analyzed in order to propose a new metric for evaluating the richness of the vocabulary used by authors of different documents by adding the semantic layer as a further characterization.
Set of contains words obtained from the preprocessing	{Vocabulary, richness, metrics, are, in, depth, analyzed, in, order, to, propose, a, new, metric, for, evaluating, the, richness, of, the, vocabulary, used, by, authors, of, different, documents, by, adding, the, semantic, layer, as, a, further, characterization}
Set of words obtained from the elimination of connection words	{Vocabulary, richness, metrics, depth, analyzed, propose, new, metric, evaluating, richness, vocabulary, used, authors, different, documents, adding, semantic, layer, further, characterization}
Set of extracted WordNet concepts	{Vocabulary, richness, metric, depth, analyze, propose, new, metric, evaluate, use, author, different, document, add, semantic, layer, further, characterization}
Metric result Type-Token	*Type – Token = 20/36 = 0.55*
Metric result ISRV	*ISRV = 0.26*
Compared analyze	The proposed metric, *ISRV*, weighs the result obtained by Type-Token metric in the sense of semantic similarity. Even if reducing words to concepts identified in the text WordNet unique value ratio is 0.55 (55%) is not considered the component of the semantic approach. In the analyzed text, there are different concepts from WordNet or close in similarity with a distance value which tends to 0. Thus, expressed in metric ISRV more realistic vocabulary richness found in a text document or analyzed fragment text.

Extending the analysis of the wealth of vocabulary and semantic distance between concepts with the time evolution of this characteristic oriented authors, the defining trend for this indicator.

4 Time Evolution Analysis of the Proposed Vocabulary Richness Metric

Context of analysis is given by an initial set composed of documents drawn up by a specific author for doing analysis and will register the proposed metric values ISRV for each document. This set is sorted chronologically, in order to generate a time series. Noting with D, the initial set of analyzed documents, where $D = \{D_1, D_2, \dots, D_i, \dots D_{nd}\}$ and nd is the cardinality of this set, is calculated the metric values set of ISRV applied in each document. It is noted that the documents in the set D are sorted chronologically. In this way, the time series is obtained $ISRV = \{ISRV_1, ISRV_2, \dots, ISRV_i, \dots, ISRV_{nd}\}$, with cardinality equal to the initial set of words.

The analysis aims to identify trends that you have the indicator that measures the semantic

DOI: 10.12948/issn14531305/20.3.2016.04

richness of the vocabulary used by the author, during the analyzed time series. Where there are several documents written by the author during the same year, the indicator ISRV for the year is calculated as the arithmetic average of the indicator values ISRV recorded in the documents of the equal years.

To evaluate the trend indicator authoring plays, is defined time series using three methods of estimating the trend:

- *absolute mean change method* implies a linear dependencies form an arithmetic progression in which each term of the series is formed from the initial period, first term in terms of time, and adding an algebraic delay multiplied by the mean absolute change; this method is suitable in the context of first-degree linear dependencies;

$$\overline{ISRV_{i1}} = ISRV_1 + \overline{\Delta} * (t-1), \forall i = \overline{1, nd}$$

- *average index method* It requires an exponential dependence of the shape of a geometric progression in which each term of the series is made from the original deadline by multiplying it by the average index of dynamic exponentially; this method is unsuitable for an exponential dependence between the indicator ISRV and series of time periods;

$$\overline{ISRV_{i2}} = ISRV_1 + (\overline{I})^{i-1} * (t-1), \forall i = \overline{1, nd}$$

- *linear regression method* is the only method proposed in the present example for analysis of the type of analytical methods and involves estimating an equation of first degree estimation carried out using the method of least squares; form the trend is given by:

$$\overline{ISRV_{i3}} = \alpha + \beta * i, \forall i = \overline{1, nd}$$

Choosing the best method for approximating the trend which it has ISRV indicator over time involves comparing the sum of squared errors caused by the three estimation proposed methods:

$$\begin{cases} S_j = \sum_{i=1}^{nd} (ISRV_i - \overline{ISRV_{ij}})^2, \forall j = \overline{1,3} \\ S^* = \max_{j=1,3} S_j \end{cases}$$

For the detailed analysis of the proposed ISRV indicator values are extracted for an author over a period of 13 years, *nd=13*. Values are given by the set:

$$ISRV = \{0.35 \quad 0.37 \quad 0.40 \quad 0.29 \quad 0.50 \quad 0.45 \quad 0.47 \quad 0.39 \quad 0.38 \quad 0.47 \quad 0.49 \quad 0.50 \quad 0.48\}$$

Figure 1 shows the evolution of ISRV indicator over 13 years.

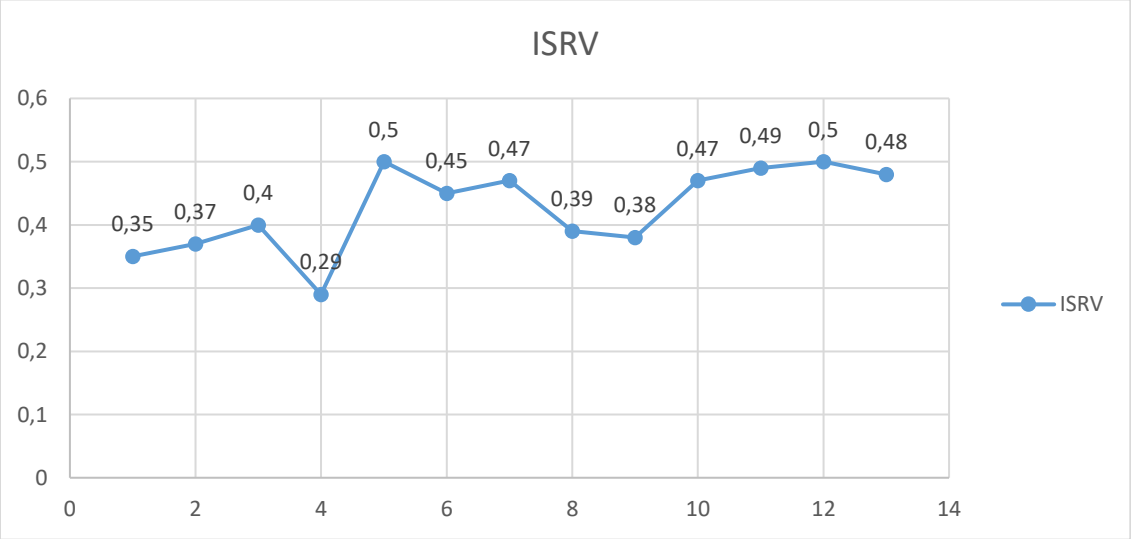

Fig. 1. Evolution of ISRV indicator

DOI: 10.12948/issn14531305/20.3.2016.04

A preliminary analysis of the resulting chart shows an increasing trend indicator ISRV value, generating an interpretation on the use of vocabulary development by increasing its level of wealth semantic analysis.

To predict developments in the next period of research, it must be running three extraction methods of trend.

Table 3 contains the calculations estimated using the first method absolute mean change method to generates series.

Table 3. Trend estimate calculations using the absolute mean change method

Year(i)	*ISRV*	$\overline{ISRV_{i1}}$	$ISRV - \overline{ISRV_{i1}}$	$(ISRV - \overline{ISRV_{i1}})^2$
1	0.35	0.350	0.000	0
2	0.37	0.381	-0.011	0.000117
3	0.4	0.422	-0.022	0.000469
4	0.29	0.323	-0.033	0.001056
5	0.5	0.543	-0.043	0.001878
6	0.45	0.504	-0.054	0.002934
7	0.47	0.535	-0.065	0.004225
8	0.39	0.466	-0.076	0.005751
9	0.38	0.467	-0.087	0.007511
10	0.47	0.568	-0.098	0.009506
11	0.49	0.598	-0.108	0.011736
12	0.5	0.619	-0.119	0.014201
13	0.48	0.610	-0.130	0.0169
			S_1	**0.076285**
$\bar{\Delta}$			**0.010833**	

Table 4 contains a series of calculations which generates estimated by using the second method, the average index method.

Table 4. Trend estimate calculations using the average index method

Year(i)	*ISRV*	$\overline{ISRV_{i1}}$	$ISRV - \overline{ISRV_{i1}}$	$(ISRV - \overline{ISRV_{i1}})^2$
1	0.35	0.350	0.000	0.00000
2	0.37	0.381	-0.011	0.00012
3	0.4	0.424	-0.024	0.00056
4	0.29	0.316	-0.026	0.00068
5	0.5	0.561	-0.061	0.00370
6	0.45	0.519	-0.069	0.00483
7	0.47	0.558	-0.088	0.00781
8	0.39	0.477	-0.087	0.00754
9	0.38	0.478	-0.098	0.00963
10	0.47	0.609	-0.139	0.01921
11	0.49	0.653	-0.163	0.02656
12	0.5	0.686	-0.186	0.03449
13	0.48	0.677	-0.197	0.03899
			S_2	**0.15411**
$\bar{I}$			**1.02913**	

DOI: 10.12948/issn14531305/20.3.2016.04

Table 5 contains the series of calculations which generates estimated using the third method, the method of linear regression.

Table 5. Trend estimate calculations using the linear regression method

	Year(i)	*ISRV*	$\overline{ISRV_{i1}}$	$ISRV - \overline{ISRV_{i1}}$	$(ISRV - \overline{ISRV_{i1}})^2$
	1	0.35	0.359230769	-0.00923	8.52071E-05
	2	0.37	0.370384615	-0.00038	1.47929E-07
	3	0.4	0.381538462	0.018462	0.000340828
	4	0.29	0.392692308	-0.10269	0.01054571
	5	0.5	0.403846154	0.096154	0.009245562
	6	0.45	0.415	0.035	0.001225
	7	0.47	0.426153846	0.043846	0.001922485
	8	0.39	0.437307692	-0.04731	0.002238018
	9	0.38	0.448461538	-0.06846	0.004686982
	10	0.47	0.459615385	0.010385	0.00010784
	11	0.49	0.470769231	0.019231	0.000369822
	12	0.5	0.481923077	0.018077	0.000326775
	13	0.48	0.493076923	-0.01308	0.000171006
∑	91	5.54		S_3	**0.0312**
	α			**0.348077**	
	β			**0.011154**	

In summary, the values obtained in this analyze is presented in Table 6 which contains the sum of the squares of errors with the equation for estimating the trend indicator ISRV obtained for the three methods of assessment.

Table 6. The sum of squared errors and the estimated trend equations for the three proposed estimation methods

Estimation method	**The sum of squared errors**	**Estimated trend equation**
Absolute mean change method	$S_1 = 0.07$	$\overline{ISRV_{i1}} = 0.35 + 0.01 * (t-1)$
Average index method	$S_2 = 0.15$	$\overline{ISRV_{i2}} = 0.35 * (1.02)^{i-1}$
Linear regression method	$\boldsymbol{S_3 = 0.03}$	$\overline{ISRV_{i3}} = 0.34 + 0.011 * i$

As for the method using linear regression has been obtained the lowest summation of squared errors, for estimating the trend equation was chosen the equation:

$$\overline{ISRV_{i3}} = 0.34 + 0.011 * i \, .$$

Figure 2 contains chart trend estimated in the estimation using linear regression. There is an increasing evolution of the trend, with 0.011 percentage points at a time to another. The interpretation is given by an expansion of the area using concepts extracted from the documents written by author.

DOI: 10.12948/issn14531305/20.3.2016.04

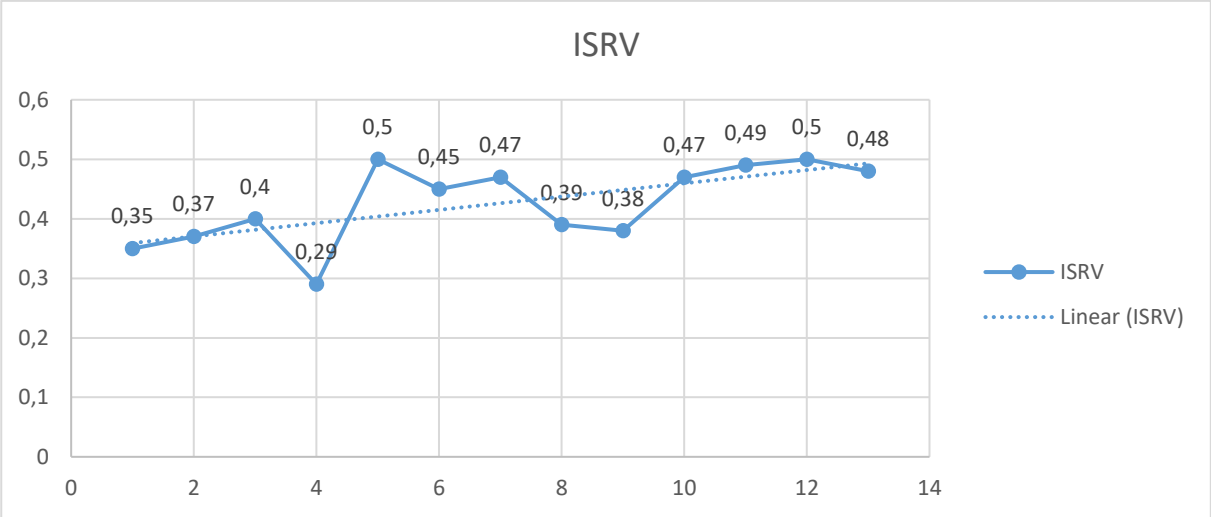

Fig. Error! No text of specified style in document.**.** Chart of ISRV indicator evolution using linear regression

The advantages of this method are that the proposed metric for assessing the richness of the vocabulary does not depend on the fields that are treated in the documents reviewed, but on the semantic distance between unique concepts identified in those documents. Adding time analysis component, resulting in a possible estimate of future works written by authors who are known previously written works in terms of time.

5 Conclusions

Transforming the vocabulary richness indicator into a semantic one adds a new layer of analysis within the general intrinsic plagiarism detection methods. First step in detecting the plagiarism is defining the author's mark within its written papers, that leads to a parts of documents analysis of similarity. Minimizing the set of phrases considered to be plagiarized, the entire process of plagiarism detection is diminished, using as input data for the next step, the external plagiarism, only those parts of documents that are considered to be different in terms of author mark analysis.

The present proposed vocabulary richness metric using semantic layer does not depend on the main subjects of the documents written by a particular author, thereby removing the subject dependency. In particular, multiple authors tend to expand their research into different domains. Using this expansion of subject non-dependency, a time evolution analysis is conducted, making possible a forecast for future time works. The present paper addresses only the problem of English written documents due to the use of WordNet lexical ontology for extracting the semantic distance and type-tokens found within the analyzed author's works. Future work are directed to the use of a Romanian lexical ontology for extracting the authors' marks within Romanian written documents.

References

[1] J. Grieve "Quantitative Authorship Attribution: An Evaluation of Techniques", Literary and Linquistic Computing, 2007, Vol. 22, No. 3, pp. 251-270

[2] D.L. Hoover "Another Perspective on Vocabulary Richness", Computers and the Humanities, 2003, Vol. 37, pp. 151-178

[3] S.M. Eissen, B. Stein & M. Kulig, 2007 "Plagiarism Detection Without Reference Collections", Advances in Data Analysis Studies in Classification, Data Analysis and Knowledge Organization, pp. 359-366

[4] E. Stamatatos & M. Koppel, 2011 "Plagiarism and authorship analysis: introduction to the special issue", Lang Resources & Evaluation, vol. 45, no. 1 , pp. 1-4

[5] E. Stamatatos, 2008 "Author identification: Using text sampling to handle the class imbalance problem", Inf. Process Manage, vol. 44, pp. 790-799

DOI: 10.12948/issn14531305/20.3.2016.04

[6] E. Stamatatos, 2009 "Intrinsic plagiarism detection using character n-gram profiles", Proceedings SEPLN, Donostia, Spain, pp. 38-46
[7] E. Stamatatos, 2009 „A survey of modern authorship attribution methods", Journal American Society Information Science Technology, vol. 60, pp. 538-556
[8] D. Yang & D.M.W. Powers, 2005 „Measuring Semantic Similarity in the Taxonomy of WordNet", 28th Australasian Computer Science Conference, Newcastle, Australia, pp. 315-322
[9] G, Tambouraizis, S. Markantonatou, N. Hairetakis, M. Vassiliou, G. Carayannis & D. Tambouratzis, 2004 „Discriminating the registers and styles in the modern greek language – Part I: Diglossia in styling analysis", Literature Linguistic Comput., vol. 19, no. 2, pp. 197-220
[10] L. Seaward & S. Matwin, 2009 "Intrinsic plagiarism detection using complexity analysis", Proceedings SEPLN, Donostia, Spain, pp. 56-61

Mădălina ZURINI is currently a teaching assistant in the field of Economic Informatics. She graduated the Faculty of Cybernetics, Statistics and Economic Informatics (2008) and a master in Computer Science in 2010. In 2013 she defended her PhD research with the title "*Spatial representations and knowledge processing using ontologies*". She published more than 20 articles in collaboration or as single author. Her fields of interest are data classification, artificial intelligence, data quality, algorithm analysis and optimizations.

Alin ZAMFIROIU has graduated the Faculty of Cybernetics, Statistics and Economic Informatics in 2009. In 2011 he has graduated the Economic Informatics Master program organized by the Bucharest University of Economic Studies and in 2014 he finished his PhD research in Economic Informatics at the Bucharest University of Economic Studies. Currently he works like a Senior Researcher at "National Institute for Research & Development in Informatics, Bucharest". He has published as author and co-author of journal articles and scientific presentations at conferences.

DOI: 10.12948/issn14531305/20.3.2016.04