Stylometry Metrics Selection for Creating a Model for Evaluating the Writing Style of Authors According to Their Cultural Orientation

Mădălina ZURINI
Academy of Economic Studies, Bucharest, Romania
madalina.zurini@csie.ase.ro

The present paper starts from a short introduction of the major aspects debated regarding plagiarism and author identification, along with the principles that are at the base of forming the property rights laws within the European community and the Anglo-American one. Regardless of the community involved, plagiarism is a form of using others research, as it is or modified, and presenting it as a personal creation. The terms of creativity and plagiarism are described in an antithesis analysis, reaching to the concept of originality, defined as a property that a creative research paper has when the ideas presented within in are different from the ones already published by different authors. A metric is implemented in order to obtain a measurable value in determining the level of originality of a paper. The main ways of testing a paper of plagiarism, intrinsic and external analysis, are described for choosing the proper methodology for determining originality of scientific papers. The research leads to the stylometric analysis, a field found at the crossroad of plagiarism, originality and author identification. This stylometric analysis is done within the intrinsic plagiarism detection and is formed on the bases of a number of metrics that describe unique a writing style of a specific author. The testing platform implies using a set of research papers written by European authors and extracting the values of eight writing style metrics. A clustering is applied and the best combination of metrics is resulted.

Keywords: *Stylometry, Plagiarism, Metrics, Clustering analysis, Cultural orientation*

1 Introduction

Researches on the intellectual property rights face determining the level of originality of a research paper, in contract to the action of plagiarism which is defined as the full or partial ownership of ideas, expressions, methods or procedures and their presentation as a personal creation. In the Anglo-American laws, the economic considerations and those that refer to public politics are prevailed in the elaboration and development of the property rights laws while, in the European point of view, the moral and civil arguments based the elaboration of the same laws.

The legislative framework does not resolve identifying plagiarism and level of originality of a scientific work. The present paper aims to apply the legislative property rights in the context of publishing scientific research papers.

In practice, there are different types of plagiarism, the most common being: copy-paste, paraphrase, plagiarism through translation in different languages, artistic plagiarism, ideas plagiarism, source code and not using the proper citations. Article [1] presents the fact that plagiarism through paraphrase is analyzed, reaching to a classification of the major known types, along with a testing using software detection of plagiarism at the level of percentage of correctness by identifying the paraphrase within a text document.

The present paper is consisted in five chapters, starting from a short introduction of the major aspects debated, along with the principles that are at the base of forming the property rights laws within the European community and the Anglo-American one. Regardless of the community involved, plagiarism is a form of using others research, as it is or modified, and presenting it as a personal creation.

Chapter 2 describes the terms of creativity and plagiarism in an antithesis analysis, reaching to the concept of originality, defined as a property that a creative research paper has when the ideas presented within in are

DOI: 10.12948/issn14531305/19.3.2015.10

different from the ones already published by different authors. A metric is implemented in order to obtain a measurable value in determining the level of originality of a paper. The main ways of testing a paper of plagiarism, intrinsic and external analysis, are described for choosing the proper methodology for determining originality of scientific papers. The research leads to the stylometric analysis within the third chapter, a field found at the crossroad of plagiarism, originality and author identification. This stylometric analysis is done within the intrinsic plagiarism detection and is formed on the bases of a number of metrics that describe unique a writing style of a specific author.

Within the fourth chapter, eight stylometry metrics are extracted from a number of scientific research papers in order to obtain the best combination that describes best the writing style of an author. For that, Weka tool along with the integration of WordNet lexical ontology analysis are used, obtaining a set of four metrics that can further describe the writing style of an author according to its cultural orientation. Conclusions are highlighted in the fifth chapter along with directions for future research.

2 Creativity and Plagiarism Analysis

Creativity, seen as a form of originality, represents the characteristic of adding something new, original and appropriate to reality, defining the novelty and originality. For that, in order to analyze the level of originality of a scientific paper, it needs to create an antithesis between this component of creativity and the plagiarism one.

Starting from the objects used within the present research, scientific research papers written by Romanian and other European authors, the component of semantic phase is defined as a compact component within a paper, formed out of one or more adjacent phases, which is significantly different from the semantic phases prior or subsequent to it. To say that a work is original is similar to the result of the evaluation of a paper in terms of plagiarism.

IEO, Indicator for Originality Evaluation, is defined as being the ratio between the total number of original semantic phases reported to the total number of semantic phases found within the analysed paper.

$$IEO = \frac{nr_{original\ phrases}}{nfs}$$

where:

- $nr_{orignal\ phrases}$ represents the total number of original semantic phrases found within the analyzed paper and is equal to $Card\left(\{fs_i, \forall\, i = \overline{1, nfs}, fs_i \neq fs_{i-1}, fs_i \neq fs_{i+1}\ , fp(fs_i) < \varepsilon\}\right)$;
- nfs represents the total number of semantic phrases found within the paper and is equal to $Card\left(\{fs_i, \forall\, i = \overline{1, nfs}, fs_i \neq fs_{i-1}, fs_i \neq fs_{i+1}\ \}\right)$;
- $fp(x)$ represents the function for evaluation of the degree of plagiarism found within the semantic phrases, having values in the rage of [0; 1];
- ε represents the maximum percentage rate accepted from the plagiarism evaluation function point of view, $\varepsilon \in [0; 1]$.

This metric is closely related to the proposed definition within [15] regarding originality. Copyright law emphasizes that "originality" fundamentally mean that a work that comes from the inspiration of the author and was not copied from another source. Hence, "original" is used in the sense of the original in order to identify the source of the work originates.

The more a work contains fewer phrases that overlap with previous research, the more the work will have a higher degree of originality. This paper uses the concept of plagiarism not only in the narrow and very known of it, within the meaning of copied without concern right, moral and legal source text, but in a sense of the idea, the research topics that can influence research an author taking into account previous studies and similar to other authors. A work is original when treating a

DOI: 10.12948/issn14531305/19.3.2015.10

concept, art, new or existing situation in a unique manner compared to other studies.
The present approaches to identify plagiarism include evaluating by comparing two or more documents. The degree of similarity is used as a quantitative assessment of the similarity between two documents on the basis of a system of metrics. In the paper [2] it is proposed a classification of the main metrics used in plagiarism detection.
In the literature, there are two main strategies for identifying plagiarism approach, [3]:

- intrinsic, which has the aim of identifying the passages plagiarized by examining only the analyzed document, concluding if parts of the material are or not written by the same author, such models are presented in [3], [4];
- external, which involves assessing through comparison of the document with other existing documents within the database od material and identifying the pair of similar documents; multiple studies analyzes this problem, such as: [5], [6], [7] and [8].

Intrinsic plagiarism identification technique uses the writing style of an author as a basis for comparison. A template is constructed, consisting in features such as: statistics on the text, features syntax, parts of speech, or sets of words commonly used structural features of the text. Feature set is attached to a function evaluation criterion of changes over the analyzed text. The disadvantages of this method are highlighted in the case of works written by several authors.
On the other hand, external approach to plagiarism brings benefits for the purposes of comparing the document with other documents written by the same author as well as other documents from the same central area. The disadvantages are given by the exponential complexity in relation to the size of the database for comparison.

3 Stylometry Metrics in Intrinsic Plagiarism Analysis

In surveys such as those developed in [9], [10], [11], [12] and [13] the problems and ways to integrate plagiarism intrinsic referring also to stylometry are treated, the writing style of an author over its history of research or in a document unit.
In the intrinsic plagiarism, in which are considered internal parts of a document suspected for plagiarism, Table 1 summarizes the types of features analyzed, used software tools and resources involved.

Table 1. Text characteristics, software tools and resources involved in the case of intrinsic plagiarism

Characteristics	Examples	Software tools and resources
Lexical characteristics (character oriented)	Characters' frequency	-
	Character type (letter, punctuation marks)	Character dictionary
	Frequency of special characters (such as!, &, *)	Character dictionary
	The frequency range of characters of size n (fixed length)	Text splitter
	The frequency range of characters of size n (variable length)	Character selection
	Compression methods	Text compression software tools
Lexical characteristics (word oriented)	Element oriented: - The average length of words - The average length of sentences - The average length of syllables words	Tokenization, sentence spitting
	The richness of vocabulary:	Tokenization

DOI: 10.12948/issn14531305/19.3.2015.10

	- Ratio of single words and total words	
	Words frequency	Tokenization
	Word frequency function type	Tokenization, specific dictionaries
	Frequency words of size n	Tokenization
	The average frequency of words	Tokenization
	Lexical errors: - Misspellings (omitting or inserting letters) - Errors of form (use uppercase)	Tokenization, lexical error checking software tools
Syntactic characteristics	Parts of speech	Tokenization, sentence spitting, part of speech identification
	Frequency of parts of speech of size n	Tokenization, sentence spitting, part of speech identification
	Pieces of text	Tokenization, sentence spitting, part of speech identification
	The structure of phrases and sentences	Tokenization, sentence spitting, part of speech identification
	Frequency rewrite rules	Tokenization, sentence spitting, part of speech identification
	Syntax Errors: - Fragments of sentences - Wrong time	Tokenization, sentence spitting, syntactic errors checker
Semantic characteristics	Synonyms, polysemantic words	Tokenization, part of speech identification, words thesaurus
	Semantic dependencies	Tokenization, part of speech identification, words thesaurus
	Functional dependencies	Tokenization, part of speech identification, words thesaurus

Regardless of the methods for identifying plagiarism type intrinsic or external, it is important to assess what features should be considered in order to obtain more accurate results. These characteristics depend on the set of documents analyzed, the language they are written and the type of documents. This work addresses the type of papers articles. Also, the language in which the documents are written is English. To retrieve the original document from the crowd semantic component describing an author affiliation to culture, components of multidimensional data analysis are used to identify features that set the style of writing that optimize the objective function to extract cultural orientation.

4 Clustering Metrics for Creating a Model for Description of Author's Writing Style

For extracting the correct set of writing style characteristics which defines at the maximum level the lexical, semantic and cultural components of an author by using his own scientific papers, the initial set of characteristics must be defined. This initial set is the one on which different combinations are performed. In this way, the set of writing style characteristics is composed of the following elements:

- the average length of words;
- the average length of sentences, measured in number of words;
- the number of connection words with regards to the total number of words from the processed documents;

DOI: 10.12948/issn14531305/19.3.2015.10

- the usage frequency of special symbols like {,;.!?@#$%&*(){}[]};
- the richness of the Type-Token vocabulary;
- the semantic richness of the vocabulary.

Beside the initial set of six writing style characteristics, two more are defined which describe better the semantic component. The first characteristic is the contextual meanings indicator or *ISC* and second one is the weighted indicator of contextual meanings or *IPSC*, characteristics that are both determined based on the WordNet ontology.

In these way the following variables are used:

- w_i is the *i* word from the set of words found in the processed document;
- $s(w_i)$ is the contextual meaning returned for the word w_i using the Word Sense Disambiguation component available in the WordNet ontology;
- $p(s(w_i))$ is the occurrence weight assigned for the contextual meaning returned by the $s(w_i)$ for the w_i word, weight that is determined based on a training set and taken from the WordNet lexical ontology.

The *ISC* is the contextual meanings indicator, meanings that the author is using them in average in his scientific papers. The *IPSC* is the weighted indicator of contextual meanings that an author uses them in average in his scientific papers weighted with the occurrences probability of the meanings found in the WordNet ontology.

The two indicators, *ISC* and *IPSC*, complete the initial set of characteristics by integrating the usage analysis of common or particular meanings of polysemantic words. The *ISC* indicator is based on the following formula:

$$ISC = \frac{\sum_{i=1}^{n} s(w_i)}{n}$$

where:

- n is the size of the set that includes the total number of words extracted from the analysed document; this set is not reduced by eliminating the redundant words because of the possibility of using multiple meanings of one word in the same document depending on context.

On the other side, the *IPSC* indicator, includes the *ISC* but in a more improved form by integrating the occurrences probability of each contextual meaning, using the following formula:

$$IPSC = \frac{\sum_{i=1}^{n} s(w_i) \times \frac{1}{p(s(w_i))}}{n}$$

IPSC is an inversely proportional variable to the occurrences weight of contextual meanings of polysemantic words:

$$\begin{cases} p(s(w_i)) \to 1 \Rightarrow \dfrac{1}{p(s(w_i))} \to 0 \Rightarrow IPSC \to 0 \\ p(s(w_i)) \to 0 \Rightarrow \dfrac{1}{p(s(w_i))} \to \infty \Rightarrow IPSC \to \infty \end{cases}$$

The zero value for this variable means that common meanings have been used in contrast with the case when the value of this indicators tends to infinity, $IPSC \to \infty$, meaning that the author frequently uses uncommon contextual meanings of polysemantic words.

For choosing an optimal set of characteristics that would describe better the cultural affiliation of an author's scientific papers, the set of combinations between these eight characteristics of size *NC* is defined.

$$NC = C_8^1 + C_8^2 + \cdots + C_8^8 = 2^8 - 1 \; possible \; combinations$$

For choosing the optimal combination an objective function is defined which must comply with the restrictions of cluster formation in an unsupervised classification:

- minimizing the inter-cluster dispersion;
- maximizing the intra-cluster dispersion.

DOI: 10.12948/issn14531305/19.3.2015.10

The role of these two conditions is that the objects groups of each combination to be as closely packed together as possible and in the same time clearly delimited between them.
A group is that set of documents written by those authors which have the same origin country. In this regard, the clustering is made up to the level which defines the country of origin.

For extracting the cultural component, the centroid is selected, called also the average value, for each cluster in hand. In figure 1, objects A, B, C, D, E, F, G are a representation example of scientific papers, so that, after an analysis of the similarity between them to gradually form clusters.

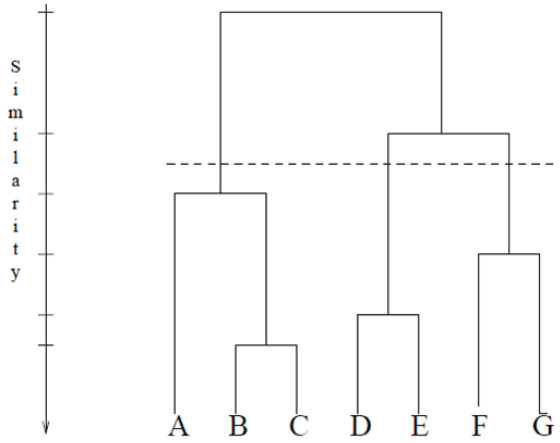

Fig. 1. Representation of object clustering using the hierarchical clustering algorithm, [14]

The dotted line in the Figure 1 is the level defined by authors belonging to countries of the same origin, thus extracting the cultural component. For the example in figure 1, the first group consists of items A, B and C, the second group of objects D and E, and the third and final group should cover the last two items: F and G.
If an analysis for a higher level grouping of countries is desired, then the hierarchical clustering algorithm stops above the level defined by the dashed line, level where the groups are actually made up of multiple groups of countries previously analyzed.
In order to identify a correlation between the analyzed characteristics, an open source application is required for converting Excel files in ARFF file type, format needed for the data mining analysis.
For creating an ARFF file using this EXCELtoARFF conversion applications, the Excel file path is chosen. An example of loading the excel file into the application in order to be converted is depicted in Figure 2.

DOI: 10.12948/issn14531305/19.3.2015.10

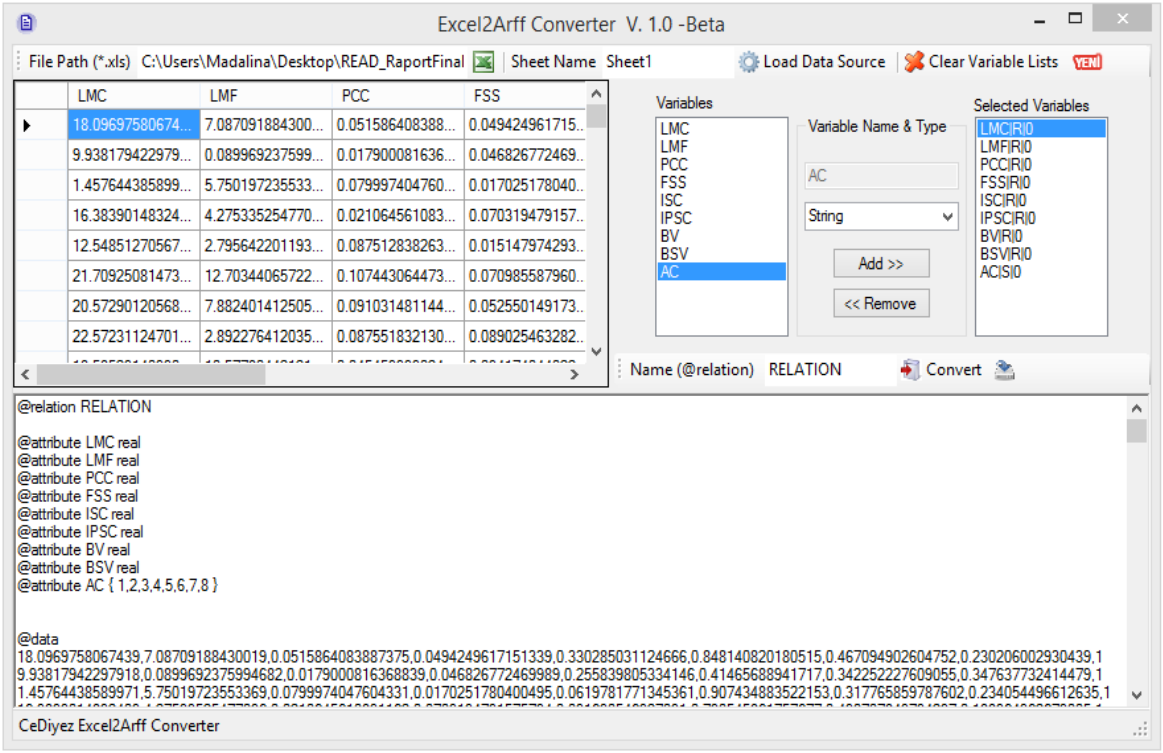

Fig. 2. Input data from Excel to Excel2Arff Converter

Table 2 contains the notations made for the variables used in the model.

Table 2. Notation for variables used within the model

Lexical characteristics	Notation	Semantic characteristics	Notation
the average length of words	LMC	contextual meanings indicator	ISC
the average length of sentences, measured in number of words	LMF	the weighted indicator of contextual meanings	IPSC
the number of connection words with regards to the total number of words from the processed documents	PCC	the richness of the Type-Token vocabulary	BV
the usage frequency of special symbols like, {, ; . !? @#$%& ∗ (){}[]}	FSS	the semantic richness of the vocabulary	BSV
cultural affiliation	AC		

The next step is to configure the file header in ARFF format it with the field leading to the right application. The list of variables is found in the left side with variable names from the Excel file, which is extracted automatically from the list the name of each column. List on the right is the list of names of the attributes that are used and integrated into the ARFF file. By selecting a variable on the left and right becomes active and selectable variable type as real variables or strings.

DOI: 10.12948/issn14531305/19.3.2015.10

In Figure 3, the variable LMC is selected to be configured by choosing real type, and automatically named 'LMC | R | O', using a standard prefix, '| R | O'. By selecting the Add button, variables are selected ARFF file. Each variable must be selected and configured for the following analysis.

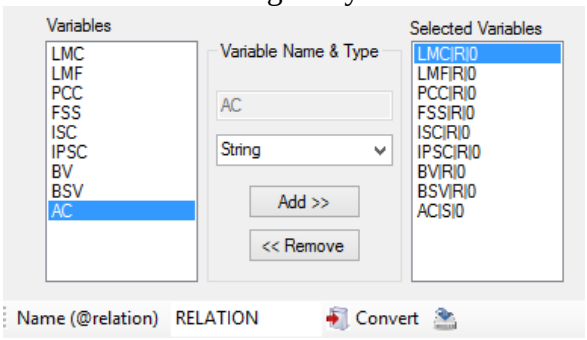

Fig. 3. Variable configuration for inserting in ARFF format

Weka, Waikato Environment for Knowledge Analysis, is a software package developed by University of Waikato, New Zealand, available at http://www.cs.waikato.ac.nz/ml/weka/. The available operations performed using the Weka open source utility are preprocessing, classification, data clustering, association rules application and selecting attributes and data visualization.

Figure 4 highlights this information after inserting the path to the file data Digital Economy Ranking.

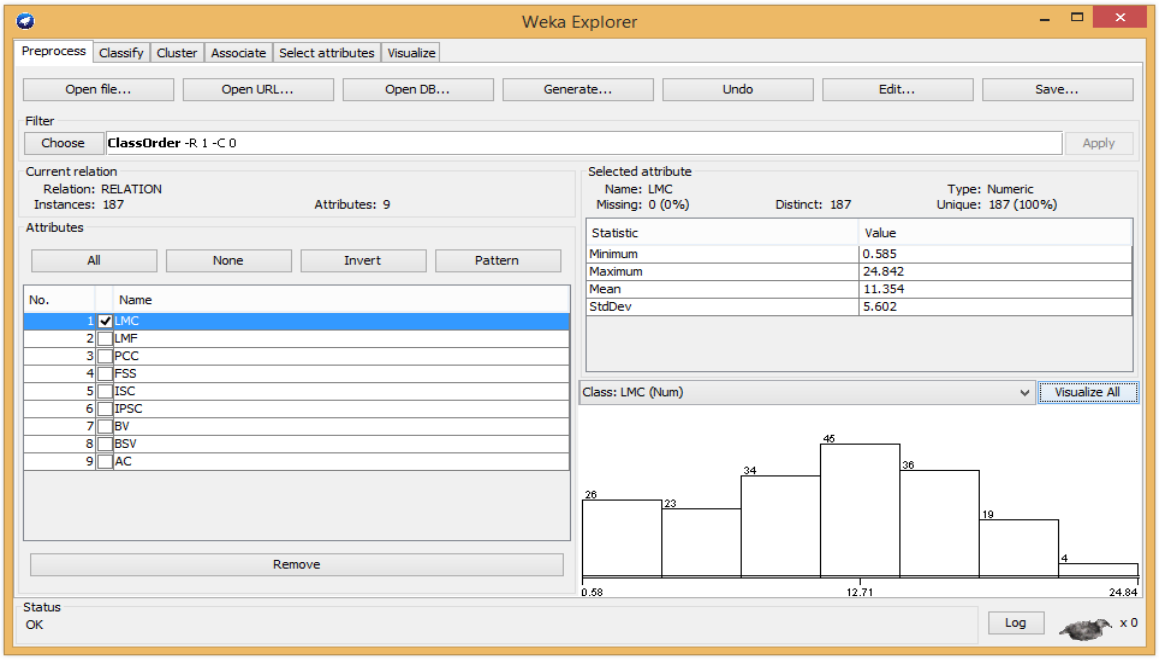

Fig. 4. Preprocessing phase

For a detailed analysis, figure 5 contains the statistical indicators calculated for each feature. ISC has a minimum value threshold of 0.001 and the maximum value of 0795. The average value for all analyzed objects is 0.354 with a standard deviation of 0.161.

DOI: 10.12948/issn14531305/19.3.2015.10

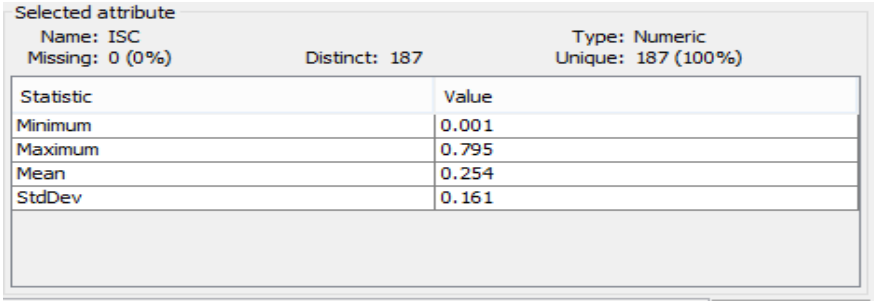

Fig. 5. Statistics upon ISC variable

Table 3 contains the values generated by Pearson correlation between any two variables in the model representation, thereby generating the correlation matrix.

Table 3. Correlation matrix for the input variables

	LMC	*LMF*	*PCC*	*FSS*	*ISC*	*IPSC*	*BV*	*BSV*	*AC*
LMC	1								
LMF	-0.09288	1							
PCC	0.019449	-0.22092	1						
FSS	0.021308	0.002204	0.190086	1					
ISC	-0.06258	0.010711	-0.07782	-0.07395	1				
IPSC	0.003019	0.02226	-0.11842	0.16367	-0.04387	1			
BV	0.111966	0.094326	-0.02442	-0.06536	0.070833	0.042996	1		
BSV	-0.01219	0.159938	-0.28752	-0.00267	-0.01054	0.000976	-0.16187	1	
AC	-0.01207	0.081905	-0.35445	0.057475	0.263434	-0.06313	-0.04487	0.237055	1

Component selection attributes in Weka is dealing with principal component analysis, analysis is performed in order to minimize duplication of information and the relationship of the courts. For the analyzed example, the eight features are transformed with the help of six new vectors used to explain the information. Transformation is generated in expression attributes of their values in the context of minimizing the loss of information, keeping the data size as small as possible. Figure 6 highlights the generated eigenvalues.

DOI: 10.12948/issn14531305/19.3.2015.10

```
Eigenvectors
 V1       V2       V3       V4       V5       V6
-0.1913   0.4351   0.1871   0.4674   0.6366   0.2337 LMC
 0.4236   0.4129  -0.1566   0.0116  -0.2545  -0.2817 LMF
-0.523    0.2491  -0.0418   0.0462  -0.3826  -0.1832 PCC
 0.0103   0.5361  -0.4515   0.2758  -0.2236   0.066  FSS
 0.3891   0.335    0.3682  -0.2554   0.2632  -0.4771 ISC
-0.0761   0.0646  -0.6762  -0.5316   0.4624   0.0324 IPSC
-0.1842   0.4167   0.3696  -0.5876  -0.1876   0.4786 BV
 0.565   -0.0276  -0.074    0.1001  -0.1238   0.6086 BSV

Ranked attributes:
 0.798   1 0.565BSV-0.523PCC+0.424LMF+0.389ISC-0.191LMC...
 0.627   2 0.536FSS+0.435LMC+0.417BV+0.413LMF+0.335ISC...
 0.489   3 -0.676IPSC-0.451FSS+0.37 BV+0.368ISC+0.187LMC...
 0.37    4 -0.588BV-0.532IPSC+0.467LMC+0.276FSS-0.255ISC...
 0.261   5 0.637LMC+0.462IPSC-0.383PCC+0.263ISC-0.254LMF...
 0.17    6 0.609BSV+0.479BV-0.477ISC-0.282LMF+0.234LMC...

Selected attributes: 1,2,3,4,5,6 : 6
```

Fig. 6. Results of principal components analysis using a selection threshold of 80%

Figure 7 contains the bi-dimensional representation of the descriptive information upon the eight variables used within the model.

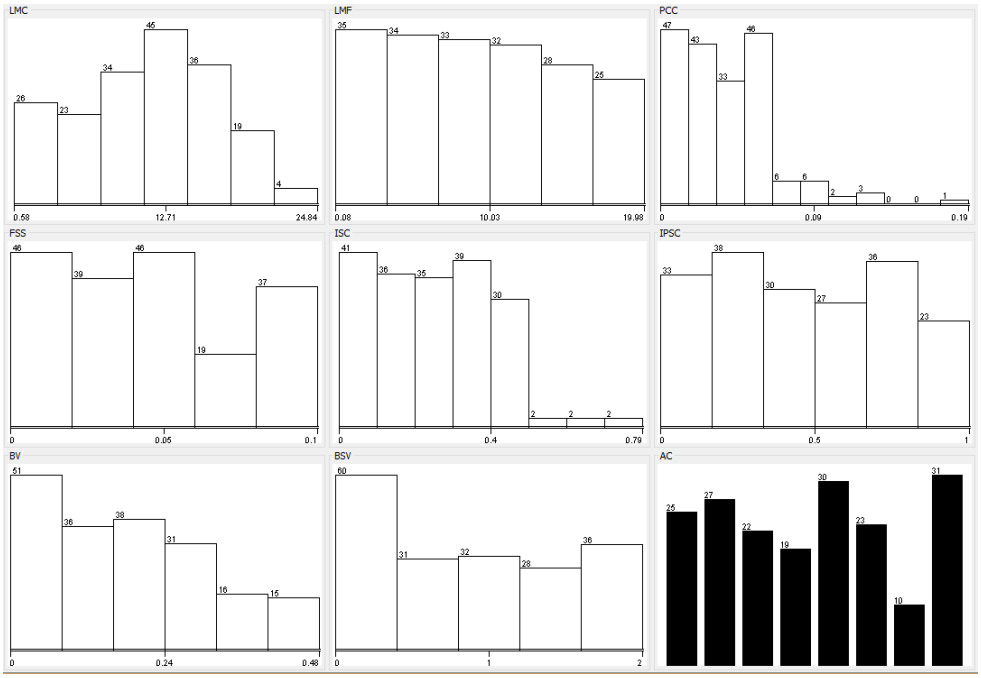

Fig. 7. Descriptive information upon the variables used in the model

To analyze which is the combination of variables that represent the right part of the cultural affiliation of papers, a first step involves applying clustering analysis using the principle kMeans to generate the final eight groups, eight countries analyzed. Figure 8 contains the result of applying the algorithm kMeans.

DOI: 10.12948/issn14531305/19.3.2015.10

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          60               32.0856 %
Incorrectly Classified Instances       127               67.9144 %
Kappa statistic                          0.2128
Mean absolute error                      0.183
Root mean squared error                  0.3152
Relative absolute error                 84.4554 %
Root relative squared error             95.7787 %
Total Number of Instances              187

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.84      0.08       0.618     0.84      0.712      0.975    1
                 0.444     0.181      0.293     0.444     0.353      0.746    2
                 0.227     0.085      0.263     0.227     0.244      0.628    3
                 0.158     0.054      0.25      0.158     0.194      0.632    4
                 0.267     0.115      0.308     0.267     0.286      0.619    5
                 0.13      0.073      0.2       0.13      0.158      0.566    6
                 0         0.045      0         0         0          0.598    7
                 0.258     0.154      0.25      0.258     0.254      0.672    8
Weighted Avg.    0.321     0.108      0.297     0.321     0.302      0.688

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h   <-- classified as
 21  1  0  1  2  0  0  0 |  a = 1
  3 12  3  3  0  0  2  4 |  b = 2
  3  6  5  0  1  2  1  4 |  c = 3
  2  6  1  3  1  3  0  3 |  d = 4
  1  4  3  2  8  4  4  4 |  e = 5
  3  4  2  0  5  3  0  6 |  f = 6
  0  2  2  1  2  0  0  3 |  g = 7
  1  6  3  2  7  3  1  8 |  h = 8
```

Fig. 8. Results for the objects' clustering

The centroids' values resulted in the clustering method applied within figure 8 are highlighted in Figure 9.

```
Number of iterations: 10
Within cluster sum of squared errors: 55.25120988404185
Missing values globally replaced with mean/mode

Cluster centroids:
                        Cluster#
Attribute    Full Data          0          1          2          3          4          5          6          7
                 (187)       (35)       (17)       (27)       (25)       (20)       (23)       (27)       (13)
=============================================================================================================
LMC              11.354    11.3251    10.5853    11.4706    11.7412    14.9928    11.0408    11.3045      6.509
LMF              9.4406    13.7694      8.523    16.4861    10.2236     7.7012     5.3409     4.5175     3.0019
PCC              0.0397     0.0321     0.0307     0.0325     0.0415     0.0813      0.035     0.0389     0.0296
FSS              0.0458     0.0665     0.0211     0.0315     0.0841     0.0661      0.016     0.0231     0.0464
ISC              0.2541     0.3344     0.1696     0.3114     0.1673     0.2486     0.2042     0.2707     0.2589
IPSC             0.4761     0.4492     0.8357     0.4355     0.5586     0.4815     0.3429     0.2259     0.7515
BV               0.1884     0.1748     0.1493     0.2527     0.1059     0.3555     0.1604     0.1679     0.1369
BSV              0.8993     1.7187      0.472     0.8246     0.5034     0.3377     1.5564     0.3768     0.9553
```

Fig. 9. Centroids values for the eight formed clusters

Running all the proposed combinations, the best which describes the cultural component determined by the cultural orientation of an author to its country origin is the one represented by the set {ISC, LMC, IPSC and BSV}.

5 Conclusions

The present paper addresses the problem of integrating the semantic analysis within the cultural orientation of authors of scientific research papers. Conducting the analysis of stylometric metrics found in the literature re-

DOI: 10.12948/issn14531305/19.3.2015.10

view, six main characteristics are extracted and added within the current analysis.
The optimization is done by adding two original metrics that combine the known metrics with a superior analysis of the semantic layer of the writing style of European authors. The objective of the paper is reached by clustering the objects represented by documents written by European authors and extracting the set of characteristics that characterize the best as possible the writing style according to the cultural orientation of the authors.

Acknowledgments
„This work was financially supported through the project "Routes of academic excellence in doctoral and post-doctoral research - READ" co-financed through the European Social Fund, by Sectoral Operational Programme Human Resources Development 2007-2013, contract no POSDRU/159/1.5/S/137926."

References

[1] A.B. Cedeno, M.Vila, M.A. Marti and P. Rosso, "Plagiarism Meets Paraphrasing: Insights for the Next Generation in Automatic Plagiarism Detection", Computational Linguistics, 2013, Vol. 39, No. 4, pp. 917-947
[2] T. Lancaster and F. Culwin, "Classifications of Plagiarism Detection Engines", ITALITCS, 2005, Vol. 4, Nr. 2
[3] G. Oberreuter, G. L'Huiller, S.A. Rios and J.D. Velasquez, "Approaches for Intrinsic and External Plagiarism Detection", Notebook for PAN at CLEF, 2011 , Available online at: http://ceur-ws.org/Vol-1177/CLEF2011wn-PAN-OberreuterEt2011.pdf
[4] E. Stamatatos and M. Koppel, "Plagiarism and authorship analysis: introduction to the special issue", Lang Resources & Evaluation, 2011, vol. 45, no. 1 , pp. 1-4
[5] N. Carnahan, M. Huderle, N. Jones, C. Stephan, T. Tran and Z. Wood-Doughty, "Plagiarism Detection", Available online at: http://www.cs.carleton.edu/cs_comps/1314/dlibenno/final-results/plagcomps.pdf
[6] S. M. Alzahrani, N. Salim and A. Abraham, "Understanding Plagiarism Linguistic Patterns, Textual Features and Detection Methods", IEEE Transaction on Systems, Man and Cybernetics – Part C: Applications and Reviews, 2012, Vol. 42, No. 2
[7] S.D. Salunkhe and S.Z. Gawali, "A Plagiarism Detection Mechanism using Reinforcement Learning", International Journal of Advance Research in Computer Science and Management Studies, 2013, Vol. 1, No. 6, pp. 125-129
[8] S.M. Alzahrani, N. Salim, A. Abraham and V. Palade, "iPlag: Intelligent Plagiarism Reasoner in Scientific Publications", Information and Communication Technologies (WICT), 2011 World Congress on, Mumbai, 2011, 11-14 Dec 2011, pp. 1-6
[9] S.M. Eissen, B. Stein and M. Kulig, "Plagiarism Detection Without Reference Collections", Advances in Data Analysis Studies in Classification, Data Analysis and Knowledge Organization, 2007, pp. 359-366
[10] B. Stein, N. Lipka and P. Prettenhofer, "Intrinsic plagiarism analysis", Language Resources & Evaluation, 2010, Vol. 45, No. 1, pp. 63-82
[11] M. Zu Eissen, B. Stein and M. Kulig, "Plagiarism detection without reference collections", Advances in Data Analysis, 2007, pp. 359-366
[12] E. Stamatatos, "Author identification: Using text sampling to handle the class imbalance problem", Inf. Process Manage, 2008, vol. 44, pp. 790-799
[13] S. Benno, K. Moshe and S. Efstathios, „Plagiarism analysis, authorship identification and near-duplicate detection", Proceedings ACM SIGIR Forum PAN'07, 2007, New York, pp. 68-71
[14] Data Mining Algorithms in R/Clustering/ Hybrid Hierarchical Clustering, Available online at: https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Clustering/Hybrid_Hierarchical_Clustering
[15] M-E. Osiceanu, „Considerații privind

drepturile de proprietate intelectuală în știință, tehnică și artă sau între creație și plagiat", Available online at: http://api.ning.com/files/uPa7BpseSwF6lqvQmgiaPdijUqzZEL9nHLQzkOJht94wzdjkfub-Wxs5cGMbkITg3agVjj0s2dOhxhjn88Hy*72*M4OH2MIVb/Osiceanu_MEConsideraiiprivinddrepturiledeproprietateintelectuala_final.pdf

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.

DOI: 10.12948/issn14531305/19.3.2015.10