Statistics and Results of Ontology Based Document Processing Application

Mădălina ZURINI
Academy of Economic Studies, Bucharest, Romania
madalina.zurini@gmail.com

The application OBDP (Ontology Based Document Processing) is presented integrating the major steps of document preprocessing, representation and automatic classification and clustering. A comparative analysis is done using the results of classification using the external knowledge base WordNet lexical ontology and the classification using Naïve Bayes and kNN classifiers. Conclusions are drawn and future work is concentrated upon WordNet extending using domain analysis.
Keywords: *Document Preprocessing, WSD, Document Classification, Clustering Analysis*

1 OBDP Application for Document Representation and Automatic Processing

Ontology Based Document Processing (OBDP) application is designed and implemented in order to validate the methodology for the optimized knowledge representation and processing steps using WordNet external knowledge base. WordNet ontology is integrated at the level of semantic analysis of the documents within the database. The application is designed as an opened system for integrating new domains and journals without modifying the logic of the application. This involves training the system to recognize (learn) new categories of documents. Also, the framework developed provides reconfigurable insights of views upon the statistics about the information stored in the database.

The set of documents processed for testing OBDP application is formed out of 1600 number of articles belonging to journals placed in areas of research: Informatics, Economy and Agriculture. Each area contains both documents used for training step and documents used for testing the developed models. The articles are taken from the following journals: *Economy Informatics*, *Journal of Information Systems & Operations Management*, *Agriculture Journal* and *Economy Journal*, totaling 700 documents used in the training step and 900 for the testing phase.

Preprocessing step comprises the introduction of documents, followed by identifying the words components and link words removal. The set of words is inserted into WSD (Word Sense Disambiguation), described in [1], [2] and [3], process after which for each word a contextual sense is assigned.

Figure 1 contains the frame for introducing the original text and displaying the new generated structure using preprocessing algorithms. To eliminate link words a set of predefined words is used. Stemming process is done using Porter algorithm and lemmatization using WordNet Lematizer.

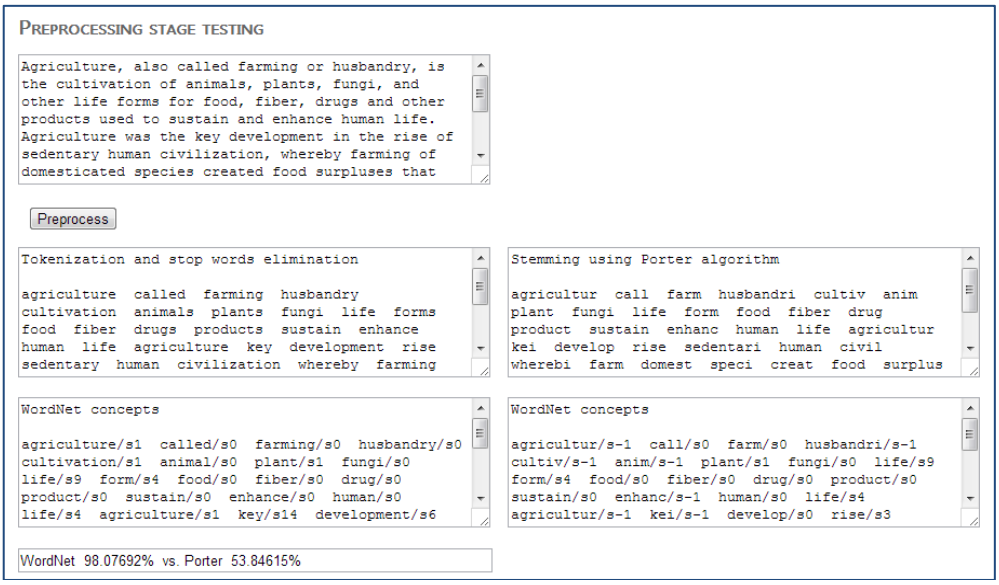

Fig. 1. Document processing's results

For a comparative analysis between Porter algorithm and WordNet Lematizer, [4] and [5], a metric based on the percentage of resulted words from WordNet is used. For the example from figure 1, WordNet Lematizer gets a score of 98.07%, compared to Porter score, only 53.84%.

After testing the two methods of stemming and lemmatizing upon the set of test, the method chosen in order to minimize the causal space by maximizing the information retained is WordNet Lematizer. After applying WordNet Lematizer, the average percentage of words' volume decrease while preserving the essence of the content of the documents is 54.09%, with a dispersion of 0.049.

After running the preprocessing step for each document within the database, the summarized results concerning the number of concepts, link words and synsets are presented in Figure 2.

The preprocessing techniques applied, (tokenization, link word removal and lemmatization), transforms the text documents into set of concepts. The resulted WordNet concepts are further integrated in the analysis using tree representation, while the concepts that are not part of WordNet are separately analyzed. For these, the co-occurrence probabilities are calculated in order to apply the similarity metrics.

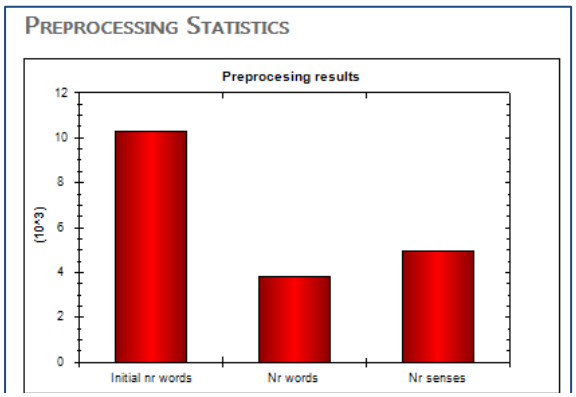

Fig. 2. Number of concepts, link words and synsets

Figure 3 contains the framework of Ontology Based Document Processing for testing the representation techniques, classification, clustering and automatic search within the set of text documents represented by articles from journals. Application components of automatic processing are highlighted.

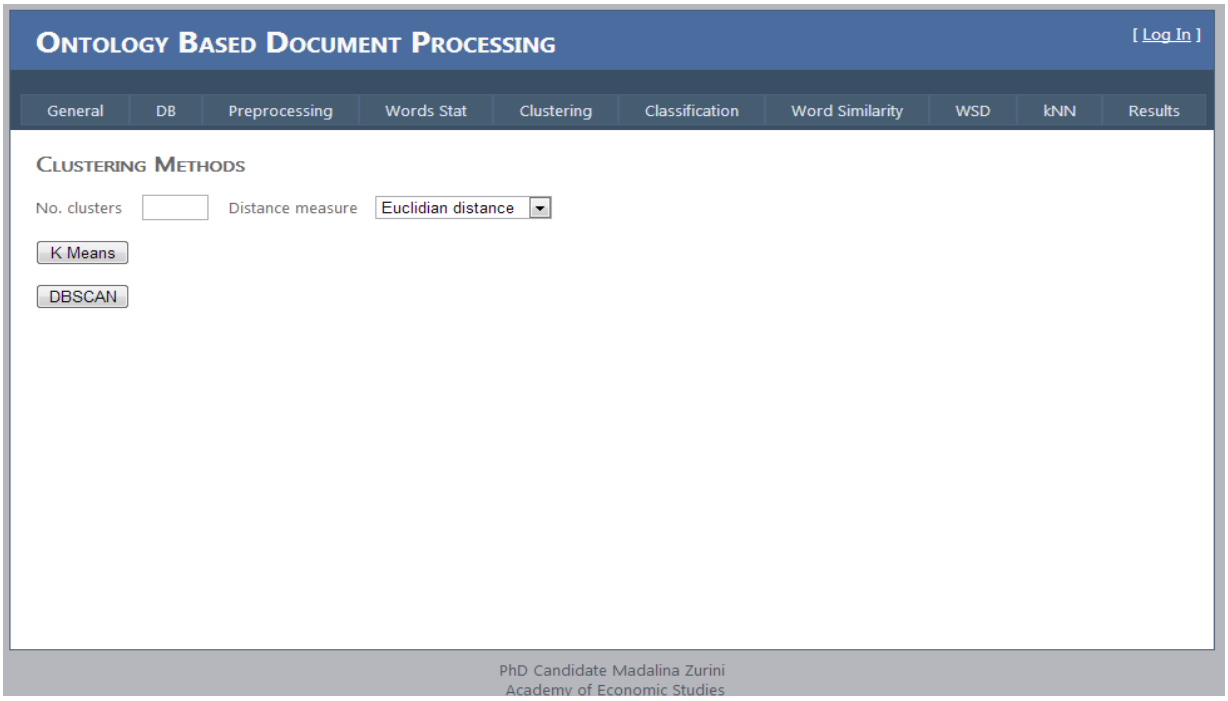

Fig. 3. OBDP Framework

OBDP components include:

- documents, words and senses representation;
- statistics upon words' grouping according to the number of appearances in the analyzed domains;
- clustering step using DBSCAN and k-Means along with the distance evaluation functions: Euclidian, Canberra, Manhattan, Cosine and WordNet Similarity;
- classification step using Naïve Bayes and kNN classifiers;
- contextual senses evaluation of polysemy words within a phrase;
- similarity measure between two concepts part of WordNet ontology;
- the main results of classification, clustering and automatic search.

2 Statistics upon Documents, Words and Senses

The purpose of multidimensional data analysis applied to the set of documents, words and senses is to extract general information available for the set of documents used as a testing base.

Figure 4 contains the structure of information stored in the database about the multitude of occurrences of words in documents

WORDS WITH SENSES FROM DB

1 2 3 4 5 6 7 8 9 10 ... >>

word	Informatics	Agriculture	Economy	UseWord
accelerating/s0	1	1	4	1
acceleration/s0	1	1	3	0
acceptable/s0	1	1	3	0
access/s0	2	2	3	0
access/s1	1	1	2	0
access/s2	1	1	2	0
access/s3	4	1	1	1
access/s4	9	1	3	1
accessibility/s0	6	1	5	1
accessible/s0	5	1	1	1

Fig. 4. Database stored information for words' senses

In order to process semantically, in the database all contextual senses are stored for the words identified in the set of documents, applying WSD algorithm, Figure 5.

For *access* word, in Figure 5 are retained the number of occurrences within each category, and in Figure 4 the occurrences of the five senses.

UNIQUE WORDS FROM DB

1 2 3 4 5 6 7 8 9 10 ... >>

word	Informatics	Agriculture	Economy	isWordNet	UseWord
acceptable	1	1	3	1	0
access	17	6	11	1	1
accessibility	6	1	5	1	1
accessible	5	1	1	1	1
accession	11	2	7	1	1
accident	1	1	2	1	0
accomplish	1	1	2	1	0
accomplished	1	1	2	1	0
accord	1	1	2	1	0
accordance	1	1	6	1	0

Fig. 5 Stored information for words

The senses' definitions of access noun available in WordNet ontology are presented in Figure 6.

Upon the set of words loaded after running preprocessing step for each document within the set of documents, the clustering analysis is done. The purpose of the analysis is to identify, based on occurrences percentages, the set of words that is common to all analyzed categories.

1. (2) entree, **access**, accession, admittance -- (the right to enter)
2. (2) **access** -- (the right to obtain or make use of or take advantage of something (as services or membership))
3. (2) **access**, approach -- (a way of entering or leaving; "he took a wrong turn on the access to the bridge")
4. **access**, access code -- (a code (a series of characters or digits) that must be entered in some way (typed or dialed or spoken) to get the use of something (a telephone line or a computer or a local area network etc.))
5. **access**, memory access -- ((computer science) the operation of reading or writing stored information)
6. **access** -- (the act of approaching or entering; "he gained access to the building")

Fig. 6. Set of senses for word access, WordNet 2.1 Browser

Figure 7 identifies the common region between Informatics and Agriculture categories.

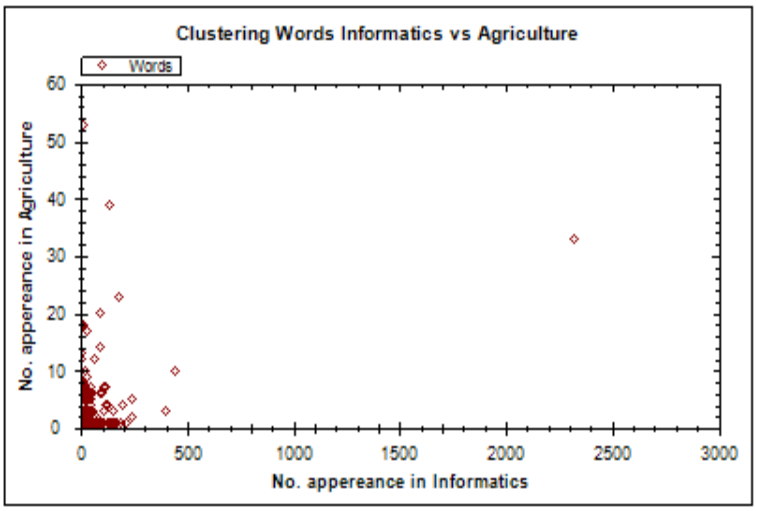

Fig. 7. Relation between words based of number of appearances in Informatics and Agriculture categories

Applying the analysis to Informatics and Economy categories, figure 8 bi-dimensional represents the words within the database. The abscissa value of each point is given by the number of appearances of the word in Informatics documents, and ordinate value measures the number of appearances of the word within Economy documents.

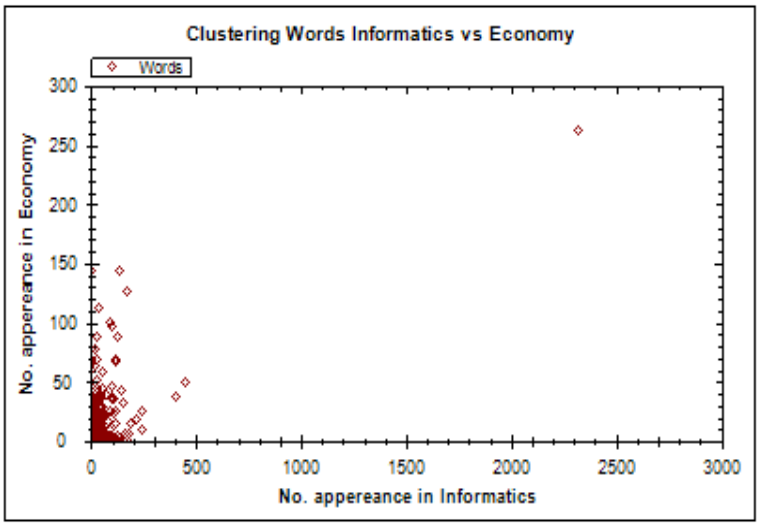

Fig. 8. Relation between words based of number of appearances in Informatics and Economy categories

Figure 9 contains the bi-dimensional representations of the words according to the appearances in the Agriculture and Economy documents.

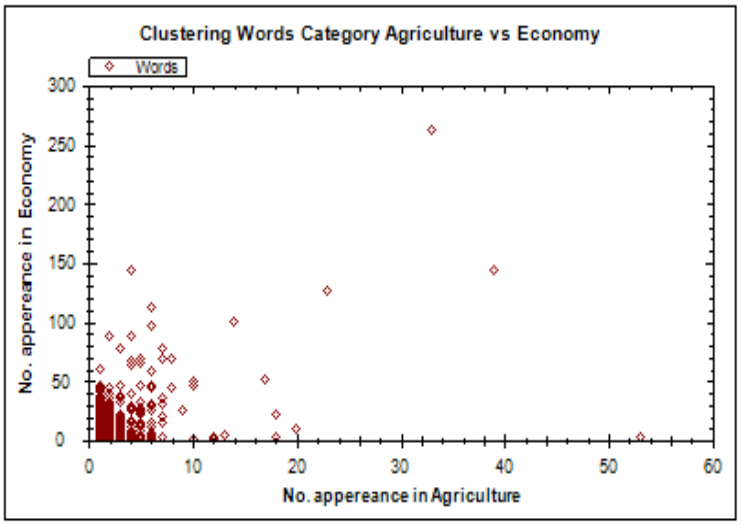

Fig. 9. Relation between words based on number of appearances in Agriculture and Economy categories

Analysis of the intersection area of the three thematic domains transposed into the set of words is based on the identification of decision threshold regarding the common or domain oriented character of each word. For that, let $w = (w_1, w_2, w_3)$ be the structure of each word formed out of the total number of appearances of the word into the domains: Informatics, Economy and Agriculture. The percentages of appearances of the word w into the domains forms the structure $w = (pw_1, pw_2, pw_3)$, where $pw_i = \frac{w_i}{\sum_{3=1}^{3} w_i}, \forall i = \overline{1,3}$ is the weight of the occurrence of the word *w* in the *i* thematic domain.

The common words are found around the bisector of the positive area of the tri-dimensional representation. If all the weights of a word are about equal, then $pw_1 = pw_2 = pw_3 = \frac{1}{3}$. Let $\varepsilon = \frac{1}{3}$, the constant used for equally separating the occurrence percentage and τ, the percentage area of the common space of the domains.

For a word to appear in the common area, then it must satisfy the relationship: $\forall i = \overline{1,3}$, $pw_i \in [\varepsilon + \tau;\ \varepsilon - \tau]$. Otherwise, the word is considered to be a domain oriented word and it will be used in the further analysis for supervised and unsupervised classification. In figure 10 it is presented the percentage of common words' evolution by adjusting the decision parameter τ, cu $\tau \in [0\%;\ 66\%]$.

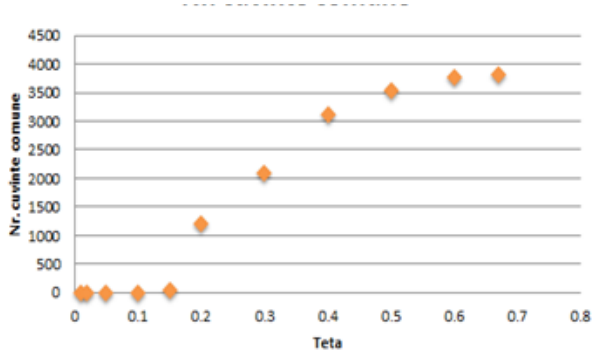

Fig. 10. Evolution graph of τ parameter and number of common words

Figure 10 contains the percentage of common words evolution from the total number of existing words within the database. If $\tau > 0.15$, the percentage of common words is grater then 20% from the total number of words. As the parameter τ increases, the area of common words is more permissive. The analysis's objective is to identify the maximum threshold of τ parameter so that the percentage of remaining words used in the automatic analysis to correctly represent the set of documents. The impact of reducing the dimension of the causal space is transposed into the percentage's evolution of correct classification.

Table 1 contains the correlation matrix between the characteristics of word variable. The strongest connection is between Informatics and Agriculture.

Table 1. Correlation matrix between numbers of appearances of words

Correlation matrix	**Informatics**	**Agriculture**	**Economy**
Informatics	1	0.4	**0.56**
Agriculture	0.4	1	0.52
Economy	**0.56**	0.52	1

Based on the correlations calculated, a hierarchy of link powers between the domains is done, identifying that the Economy domain is at the basis of Informatics and Agriculture, Figure 11.

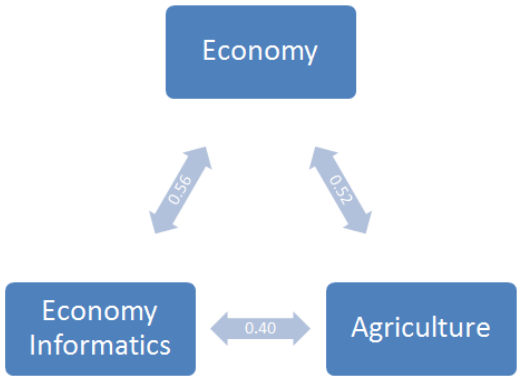

Fig. 11. Relations between categories based on common words

Table 2 contains the correlation matrix between the characteristics of senses variable, highlighting the strongest connection reached for Agriculture and Economy domains.

Table 2. Correlation matrix between numbers of appearances of senses

Correlation matrix	Informatics	Agriculture	Economy
Informatics	1	0.36	0.47
Agriculture	0.36	1	**0.51**
Economy	0.47	**0.51**	1

Based on the calculated correlations, in figure 12 a hierarchy of the connections' powers between the domains is provided. Economy stands at the basis of forming Informatics and Agriculture. The hierarchy contains the contextual semantics analysis by identifying the contextual senses of the words. By a comparative analysis between the relations of words and senses, Economy is the fundamental domain from which Informatics and Agriculture are detached.

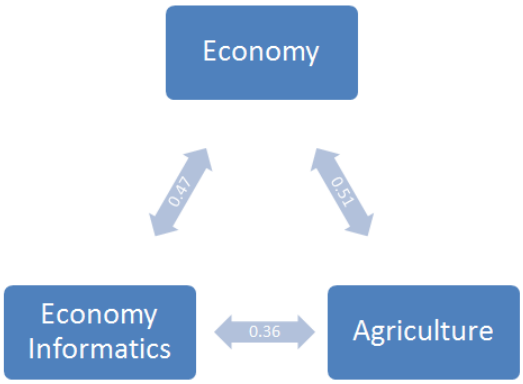

Fig. 12. Relations between categories based on common senses of words

Table 3 provides the number of words without taking into account the senses and the number of words in association with the contextual senses. The average number of senses per word recorded in the database is 1.29, for the Economy field. For this category, it is recorded the main meanings reported to the number of words.

Table 3. Statistics upon number of words within each category

Category	Nr. words	Nr. words WordNet	Nr. word-senses	Average number of senses/word
Informatics	1908		2366	1.24
Agriculture	322		410	1.27
Economy	1584		2151	1.36
Total	3814	3104	4927	**1.29**

Figure 13 contains the distribution's results of the set of documents according to the similarity function. The similarity is calculated for each of the four distances used: Euclidian, Canberra, Manhattan and Cosine. In order to be able to compare the results, normalization is applied in the interval [0; 1]. Thus, for each calculated distance, the transformation formula is used:

$$distN_k(d_i, d_j) = \frac{dist_k(d_i, d_j)}{\max\limits_{\substack{i=\overline{1,nr\_doc} \\ j=\overline{1,nr\_doc}}} dist_k(d_i, d_j)}, \forall k = \overline{1,4}, i = \overline{1, nr\_doc}, j = \overline{1, nr\_doc}$$

where:

- $distN_k(d_i, d_j)$ is the new distance between i and j documents;
- nr_doc is the set of documents size;

- k is the distance type used for evaluating the similarity between two text documents;
- $dist_k(d_i, d_j)$ is the initial distance between *i* and *j* documents.

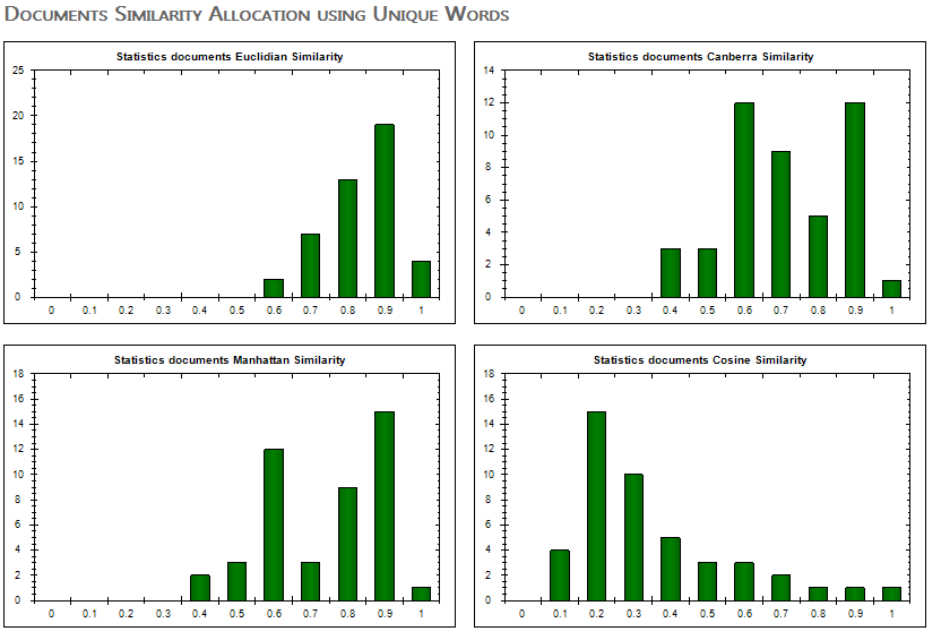

Fig. 13. Documents distributions using similarity metrics between words

Figure 14 contains the four distributions of the distances using the normalized similarity metrics within [0, 1] interval. The documents' representation method is given by Bag of Words structure having as variables the set of senses identified within the set of documents.

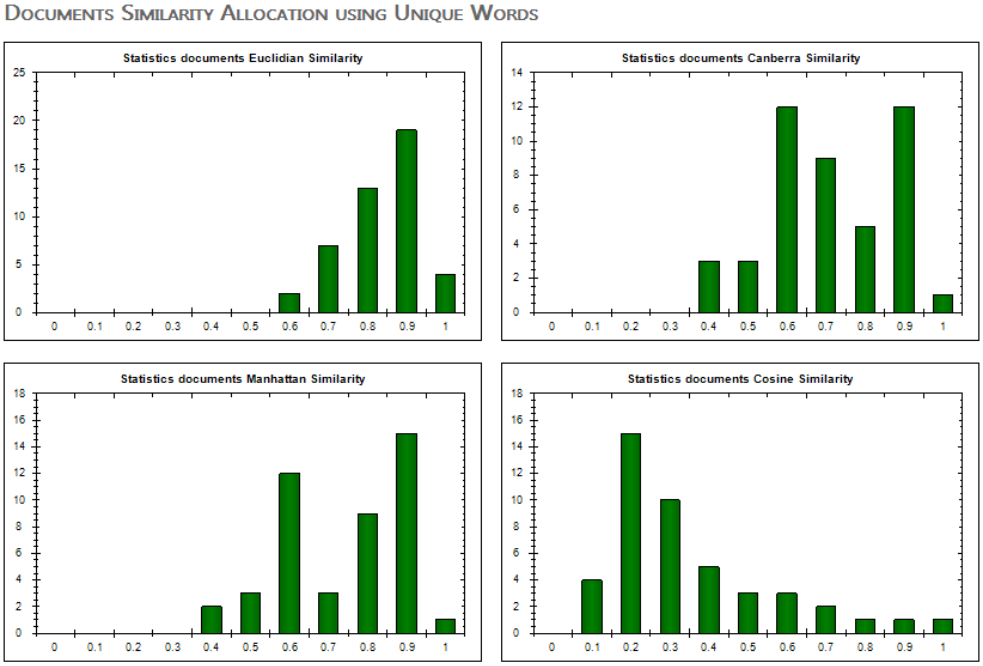

Fig. 14. Documents distributions using similarity metrics between senses

Figure 15 contains the results of the distances' distribution between each two documents using Wu & Palmer distance evaluation metric.

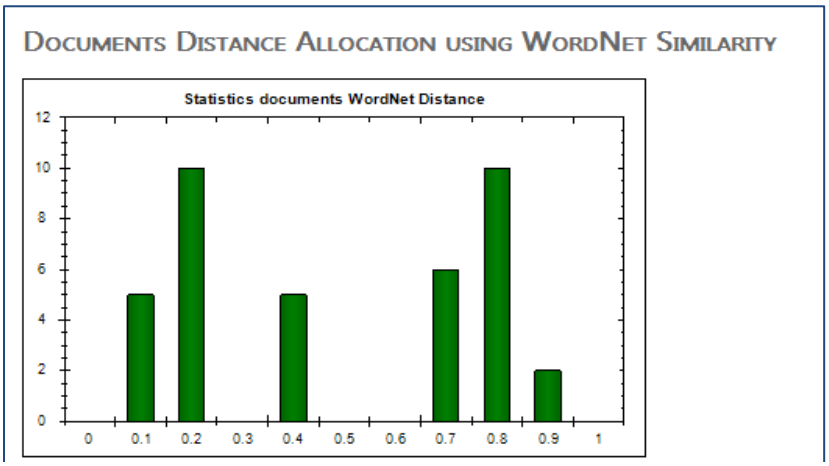

Fig. 15. Documents distribution using WordNet similarity measure

Table 4 contains the correlation matrix between the four analyzed characteristics for the given set of documents.

Table 4. Distance Correlation Matrix

Correlation Matrix	Euclidian	Canberra	Manhattan	Cosine
Euclidian	1	0.97	-0.16	0.94
Canberra	0.97	1	-0.23	**0.99**
Manhattan	-0.16	-0.23	1	-0.24
Cosine	0.94	**0.99**	-0.24	1

3 Implementation Results

Document processing for the subject based classification uses Naïve Bayes and kNN, described in [6], [7], [8] and [9], supervised classification techniques. The execution parameters of the algorithms are:

- selection of distance function used for measuring the similarity between two text documents: Euclidian, Canberra, Manhattan and Cosine distances;
- the usage of words and contextual meanings of words existing within the database.

Figure 16 contains the execution stage for training Naive Bayes and kNN classifiers.

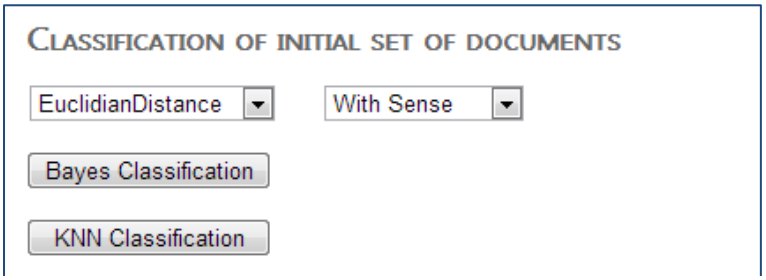

Fig. 16. The classifier's training stage in OBDP application

To test the implemented classification algorithms, Figure 17 presents the classification results of a new document inserted by a user. The input data contains

title, abstract and key words of the paper and the result is a membership class assigned by each of the two classifiers.

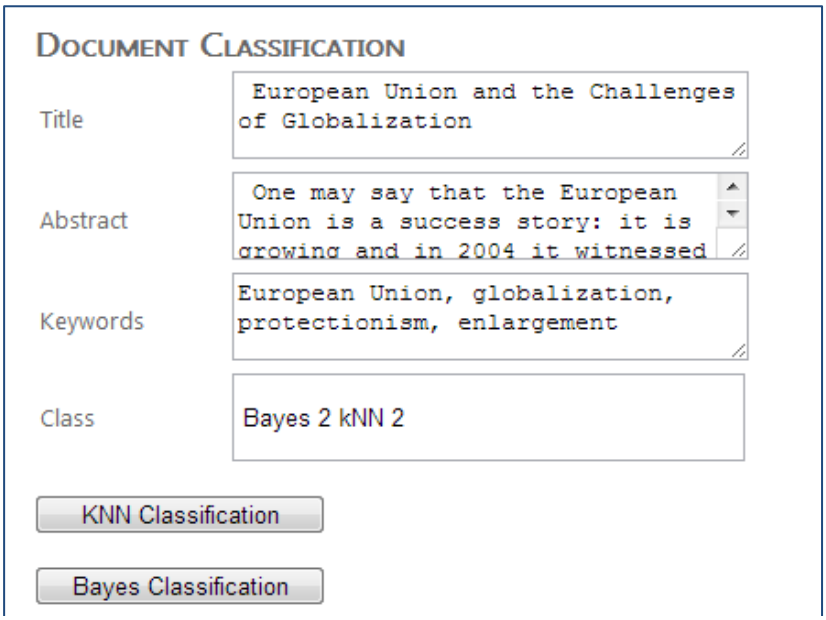

Fig. 17. Classification step of documents within OBDP application

For the classification of science articles the classification technique using kNN automated search is used of the *k* closest documents and the aggregated results by applying majority voting. Figure 18 depicts a graphical representation for the evolution of percentage of correct classification based on the size of the training set.

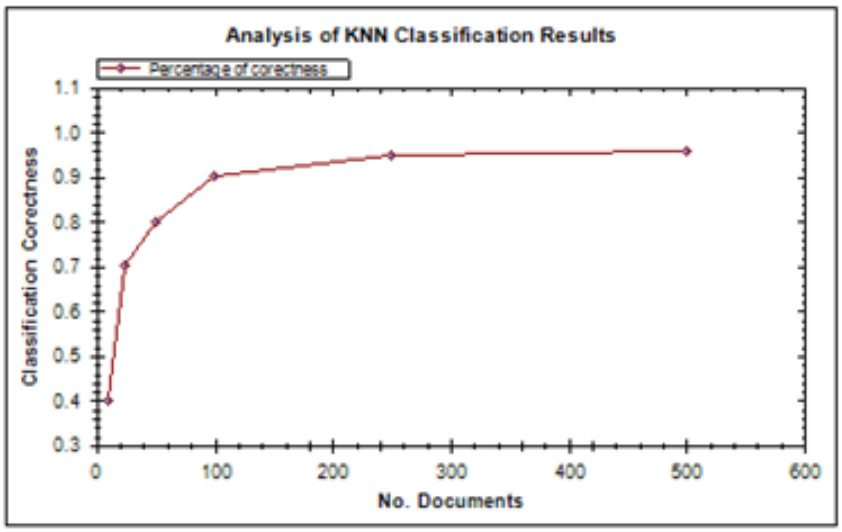

Fig. 18. kNN classification evaluation

Table 5 contains the classification values using the kNN classifier by the use of confusion matrix.

Table 5. kNN Confusion Matrix

Confusion Matrix CFM	**Classification result**		
	Informatics	**Agriculture**	**Economy**

Prior class	**Informatics**	875	10	5
	Agriculture	10	483	7
	Economy	8	2	190

Based on the matrix values presented in Table 5, the values of correctness evaluation indicators are calculated, *Precision*, *Recall* and *F Measure*, Table 6.

Table 6. Precision, Recall and F Measure values for kNN classification

Class	Precision %	Recall %	F measure %
Informatics	98.31	97.98	98.15
Agriculture	96.60	97.58	97.09
Economy	95.00	94.06	94.53
Total			**96.69**

The *F Measure* indicator for the entire classification, calculated by weighting each of the three classes with the documents percentage is 96.69%. The correctness of classification is given by the percentage of correct classified documents, *IECL*.

$$IECL = \frac{\sum_{i=1}^{nr\_clase} CFM_{ii}}{\sum_{i=1}^{nr\_clase}\sum_{j=1}^{nr\_clase} CFM_{ij}} \times 100 = \frac{875 + 483 + 190}{1600} \times 100 = 96.75\%$$

For the classification process using Naive Bayes the database is loaded with information about the probabilities of each sense for all the concepts identified in the set of documents. Based on the calculated conditional probabilities the classification process is evaluated using an available testing set like the kNN classification. Table 7 presents the results for the evaluation of Naïve Bayes classification, the confusion matrix.

Table 7. Naive Bayes Confusion Matrix

Confusion Matrix CFM		**Classification result**		
		Informatics	**Agriculture**	**Economy**
Prior class	**Informatics**	860	5	35
	Agriculture	16	477	7
	Economy	18	2	180

Based on the matrix values presented in Table 7, the values of correctness evaluation indicators are calculated, *Precision*, *Recall* and *F Measure*, Table 8.

Table 8. Precision, Recall and F Measure values for Naive Bayes classification

Class	Precision %	Recall %	F measure %
Informatics	95.56	96.20	95.88
Agriculture	95.40	98.55	96.95
Economy	90.00	81.08	85.31
Total			**93.56**

The value of F Measure indicator for the entire classification process using Naïve Bayes classifier is 93.56%. The indicator reaches its best value for the Agriculture category, of 96.95%. Table 9 presents a comparative analysis of each category of the F Measure indicator for the kNN and Naïve Bayes classifications.

Table 9. Comparative analysis between kNN and Naive Bayes based on F Measure values

Class	**n_i**	**F_Measure kNN (%)**	**F_Measure Naive Bayes (%)**
Informatics	900	**98.15**	95.88
Agriculture	500	**97.09**	96.95
Economy	200	**94.53**	85.31
Total	1600	**96.69**	93.56

kNN classification algorithm has better results than Naïve Bayes algorithm if it is evaluated for each assigned class. Naïve Bayes and kNN are used in this context as multiple class classification functions of exact association type, each object being classified in the nearest class in terms of probability.

4 Conclusions

Document processing is a technique that has grown in importance over the last period when the documents have become digital. The evolution of technology generated an exponential growth of the processed information' volume, leading to the need of integrating artificial intelligence; this meaning structuring, analysis, organization, searching and information retrieval.

Text document processing steps consist of the first phase in the introduction of the knowledge base WordNet ontology and the set of structured documents, followed by preprocessing algorithms. Within data preprocessing phase, the steps refer to dimension reduction of the documents using stemming and lemmatization. The two categories of algorithms used for stemming, with affix removing or by applying statistical algorithms lead to reducing the causal space dimension with a percentage equal to 45%.

WordNet ontology becomes part of the representation and processing of text documents steps. It is added as a database of additional description of the concepts used within the documents in order to improve the classification process.

The advantages of introducing WordNet ontology in the representation and processing flow consists in the association between each identified word with the set of existing concepts in WordNet ontology. The aggregation of the generated words using synsets conducts to a dimension reduction using BOW representation and determines the contextual sense of each polysemy word.

This process of word-concept association leads to an improvement of the classification's correctness within the developed application. For measuring the similarity between two concepts, a hierarchal representation is used based on the relations identified in WordNet. By applying WSD, the classification process is improved in terms of classification accuracy. The algorithm is tested using a set formed out of 100 phrases with 200 polysemy words. After running the algorithm for each phrase, the percentage of contextual sense classification is equal to 92%.

WSD limitations consist in the existence of WordNet only for English vocabulary along with the appropriate existing meanings. Another limitation solved is given by the number in the set of documents that are not described in the ontology. For measuring the similarity between the non-existing words in WordNet, Jaccard metric is used based on the co-occurrences of the words within the set of documents from the training stage.

Acknowledgments

This work was cofinaced from the European Social Fund through Sectoral Operational

Programme Human Resources Development 2007-2013, project number POSDRU/107/1.5/S/77213 „Ph.D. for a career in interdisciplinary economic research at the European standards".

References

[1] S. Kamali, *Some Experiments in Word Sense Disambiguation*, 2001, Available online at: https://cs.uwaterloo.ca/~s3kamali/courses/word-sense-disambiguation.pdf

[2] L. Xiaobin, S. Szpakowicz, S. Matwin, „A WordNet-based Algorithm for Word Sense Disambiguation", *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1995, pg. 1368--1374

[3] P. Resnik, „Disambiguating Noun Grouping with Respect to WordNet Senses", *Natural Language Processing Using Very Large Corpora Text, Speech and Language Technology*, Vol. 11, 1999, pg. 77-98, ISBN 978-90-481-5349-7

[4] R. Mihalcea, C. Strapparava, „Corpus-based and Knowledge-based Measures of Text Semantic Similarity", *Proceeding AAAI 06 Procedings of the 21th National Conference on Artificial Intelligence*, 2006, Vol. 1, pg. 775-780, ISBN 978-1-57735-281-5

[5] J. Xu, W. B. Croft, "Corpus-Based Stemming Using Co-occurrence of Word Variants", *ACM Transactions on Information Systems*, 1998, Vol. 16, Nr. 1, pg. 61-81

[6] D. Kolbe, Q. Zhu, S. Pramanik, „Reducing non-determinism of k-NN searching in non-ordered discrete data space", *Information Processing Letters*, 2010, pp. 420-423

[7] Y. S. Chen, Y. P. Hung, T. F. Yen, C. S. Fuh, „Fast and versatile algorithm for nearest neighbor search based on a lower bound tree," *Pattern Recognition*, 2007, pp. 360-375

[8] E. Plaku, L. E. Kavraki, „Distributed computation of the knn graph for large high-dimensional point sets", *Journal of Parallel Distributed Computation*, 2007, pp. 346-359

[9] V. F. Lopez, F. Prieta, M. Ogihara, D. D. Wong, „A model for multi-label classification and ranking of learning objects", *Expert Systems with Applications*, 2012, pp. 8878-8884

Mădălina ZURINI is currently a PhD candidate in the field of Economic Informatics. She graduated the Faculty of Cybernetics, Statistics and Economic Informatics (2008) and a master in Computer Science, having her dissertation given in *Implications of Bayesian classifications for optimizing spam filters* (2010). She is also engaged in Pedagogical Program as part of the Department of Pedagogical Studies. Her fields of interest are data classification, artificial intelligence, data quality, algorithm analysis and optimizations. She wants to pursue a pedagogical career.