Professor Ion LUNGU, PhD
The Bucharest Academy of Economic Studies
Professor Maria Teresa PAZIENZA, PhD
University of Rome, Italy
Alexandra TUDORACHE, PhD Candidate
The Bucharest Academy of Economic Studies

ORGANIC TOPIC RECOGNITION IN ONLINE DOCUMENTS

Abstract. *In this paper we describe our study on organic topic recognition and knowledge extraction from online documents. Firstly we will introduce the current state of information extraction with ontology support then we will present our approach, corpus, theoretical background, algorithm, experiments and research goals. At the end we will draw some conclusions and show some future research directions.*

Keywords: knowledge management, knowledge extraction, document classification**,** *support vector machines, Sequential Minimal Optimization, organic agriculture, AGROVOC ontology.*

JEL Classification: C10, C31, D80, Q16

1. Introduction

Nowadays people are confronting with a world food crisis not only from a quantitative point of view but also from a qualitative one. Organic farming is not a mere choice any more but a necessity. For farmers, to set up successful businesses, information access in a consistent and usable is the key.[2] They need to know what works and what doesn't work and in witch conditions of soil, climate: the so called best practices as well as to be aware of standards world wide. Simply said they need to have access to specific and structured knowledge. Knowledge comes in many forms especially as text: books, brochures, web documents, forums, questions and answers and s.o. For many of us Internet seems to be the most handy choice when searching for information but often people that need such information are not using computers and Internet on a day to day basis and don't have the necessary skills to obtain useful information from traditional search engines. Furthermore information in the WWW is written by millions of authors with different languages, ways of expressing and s.o. Also current search engines are not focused enough and they are not capable to structure and aggregate information. As a result often people can't find the information they need.

After a thorough analysis of various documents' features we found out that documents that contain information about organic agriculture and classic agriculture share similar concepts. As widely discussed in the next section a document is every piece of text that contains information in the agriculture domain.

Moreover there is not a well established corpus for testing so we had to manually

build a resource that is not homogeneous as needed.

Furthermore to overcome or at least minimize corpus problems we needed some external knowledge. We were not able to find any specific thesaurus on organic agriculture but for the general notions of agriculture we obtained the "AGROVOC" thesaurus as well as the Ontology build on top of it from the "Food and Agriculture Organization" - further referred as FAO.

The ontology will be used to disambiguate phrase and sentence domains and to obtain a set of better structured data to work with a document classifier.

The paper is structured in more section the first one presents some definitions and acronyms that will be used, then the current state of document classification with the support of Ontologies. Furthermore, it will be presented a detailed view of our model including document representation, classification models, evaluation measures and our algorithm.

Then we will widely discuss our experimental results and in the end we will explain our conclusions and further work directions.

2.Core Definitions and Acronyms

In this section, we will introduce some notions that will be used further on.

In our perspective o document is a text unit. It can be a short fragment of text or a phrase. The *corpus* is build from documents, each unit being stored in one text file. The corpus contains only documents that carry information about crop management.

Organic agriculture documents are text units that describe crop management techniques and best practices for organic agriculture like fertilizers, how to water and other cropping methods.

Standard agriculture documents are text units that contain information about the best practices and methods applicable for the classic intensive agriculture production.

AGROVOC is FAO's Agriculture Multilingual Thesaurus and covers the terminology of agriculture, forestry, fisheries, food and environment.

AOS/CS is the Agricultural Ontology Service/ Concept Server developed by FAO. AOS/CS was projected as a tool oriented to structure and standardize agricultural terminology in multiple languages. AOS uses AGROVOC as thesaurus to build the agriculture ontology. [17]

Other Agricultural thesauri are the CAB Thesaurus and the thesaurus of the National Agricultural Library in the United States.

3.Current State of Document Classification with Ontologies Support

In the previous section we introduced some notions and acronyms that will be used in this paper. In this section firstly we will introduce several methods for document classification and also some mixed methods that use Ontologies support.

For document classification we will use classic supervised learning techniques (e.g. Support Vector Machines – further referred as SVM, Naive Bayes or Maximum

Entropy). In this document we will focus on the first two methods.

In supervised learning we also need a training corpus that is build by manually labeling inputs. In the next section we will present in detail our corpus and how it was build.

Shortly the SVM algorithm divides documents represented as vectors by a hyperplane. Each word in the corpus becomes a dimension, and a vector represents some measure of the number of occurrences of the respective words in the document. (E.g. Frequency or term frequency per inverse document frequency)

Also to reduce dimensionality often are used stemming algorithms or other techniques such as lexical relations to collapse the features with the same meaning[14].

Joachims, T. [5, 6] and others [1] demonstrated that SVM has a good accuracy and works better than other algorithms. SVM works better in high-dimensional vector space (where there are many words in the vocabulary). Furthermore each document uses only a small subset of the corpus vocabulary. SVM is well suited for sparse document vectors but significant for the domain. (E.g. words that are present in a document are important). [4]

Furthermore we will present some mixed methods that use domain Ontologies as support for the classifiers. Ontologies can help people to retrieve documents that are similar or related.

Nagarajan, M. and al. have shown that by altering term vectors with the aid of semantic relations defined in a domain ontology precision and recall were greatly improved. [12]

An example of ontology enhanced information retrieval is *Textpresso* text-mining engine. With the support of a domain ontology to tag documents and enhance search by semantic queries search efficiency was improved significantly. [11]

Kwang Mong Sim developed an information filtering agent that assists users in identifying out-of-context web pages and rating the relevance of web pages. Experimental results shown that this way the ranking web pages was close to human ranking in many cases. [7]

1 Our Model

Firstly we will introduce some basic notions about the document representation, corpus, classification model, evaluation and measures we will be using. Furthermore we will be outlining the main steps of our approach including testing and training models of our algorithm.

1.1 Document representation

We will be representing documents as word or feature vectors. Firstly we will build the corpus vocabulary and then compute the tf*idf (term frequency * inverse document frequency) measure for each feature in each document.

$$tf_{t,d} = \frac{n_{t,d}}{\sum_{t} n_{t,d}} \quad (1)$$

where:

$\mathbf{n_{t,d}}$ is the number of occurrences of the considered term in document **d**

and the denominator is the number of occurrences of all terms in document.

$$idf_t = \log \frac{N}{df_t} \quad (2)$$

where:

N is the number of all documents

$\mathbf{df_t}$ is the number of documents containing the term

Tf*idf weight is lower when the term occurs fewer times or occurs in many documents. The best results are obtained with Tf*idf scoring when a word is important for a document and is not found in too many documents.

1.2 Base Corpus

Since there is no specific corpus available we had to gather and manually classify a small number of documents for testing purposes. The documents were specifically questions and answers about general crop farming in California and about organic farming in Canada.

Furthermore agriculture domain is very broad therefore is very difficult to build a homogeneous corpus. To have a well organized corpus it is necessary to have the support of a domain expert. For our experiments we acknowledge the situation and will are expecting lower results than we can obtain with a well structured corpus.

For our experiments we collected documents for two corpora. The first one is a small specific and focused corpus for both training and testing and the second one only a testing corpus.

For the first corpus we manually selected only fragments of documents containing correct information on crop management. Corpus was divided into two classes organic and classic agriculture practices. We also deleted information specific to location and maintained only crop best practices.

From this initial corpus we extracted randomly the training and testing corpus.

Often organic crop methods are presented as the opposite of classic agriculture. We dropped such phrases from training corpus to avoid biased information for the classifier. This way we obtained a nearly homogeneous training corpus.

Furthermore we build a second set of testing corpus extracted directly from web documents as parts of less specific documents. This second testing corpus was build to simulate a real life test with documents gathered from the WWW only with minor preprocessing (HTML markup removing).

We have to mention that both classes are very close as organic agriculture is a

subclass of general agriculture.

1.3. Preprocessed Corpus

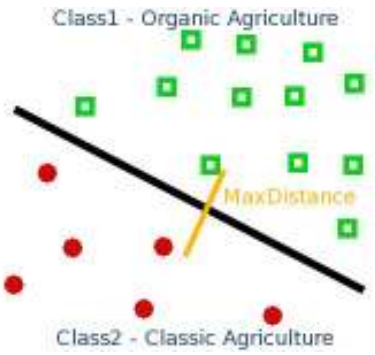

Figure 1: Linear support Vector Machines

For our experiments the base corpus described before it is transformed during the preprocessing step. Not only stopwords are removed, but since the two classes are really near and often overlapped we filter all words that are not domain specific using the AGROVOC ontology.

E.g.: "*Soybean (genetic modification) - Main article: Soybean - Soybeans are one of the "biotech food" crops that are being genetically modified, and GMO soybeans are being used in an increasing number of products. Monsanto is the world's leader in genetically modified soy for the commercial market. In 1995, Monsanto introduced "Roundup Ready" (RR) soybeans that have had a copy of a gene from the bacterium, Agrobacterium sp. strain CP4, inserted, by means of a gene gun, into its genome that allows the transgenic plant to survive being sprayed by this non-selective herbicide, glyphosate. Glyphosate, the active ingredient in Roundup, kills conventional soybeans. The bacterial gene is EPSP (= 5-enolpyruvyl shikimic acid-3-phosphate) synthase. Soybean also has a version of this gene, but the soybean version is sensitive to glyphosate, while the CP4 version is not.*"

After preprocessing same document became: "*Soybean genetic main article Soybean Soybeans biotech food crop that being genetically soybeans being increase number product world leader genetically commercial market introduce Ready soybeans that have have copy gene from bacterium Agrobacterium strain gene into genome that allow transgenic plant being this herbicide glyphosate Glyphosate active ingredient kills conventional soybeans bacterial gene acid phosphate Soybean also have version this gene soybean version glyphosate*"

1.4. Theoretical Background - Support Vector Machines Sequential Minimal Optimization

In this section we will introduce some notes about the classification models and their implementation focusing on SVM and John Platt's Sequential Minimal Optimization – further referred as SMO - algorithm for training SVM.

For our experiments we adopted WEKA - "The Waikato Environment for Knowledge Analysis" [19] and an extension of Word Vector Tool for extracting documents' word vectors. We tested with SVM implementation in Weka. Also we adopted Jena [5] for ontology support.

In 1979 Vladimir Vapnik invented a new statistical learning theory called SVM.

SVM computes a function that separates negative and positive examples with maximum margin. E.g. A linear SVM is a hyperplane that separates a set of positive examples in our case organic agriculture topics from a set of negative examples – general agriculture topics. Each word is represented by a point in feature space. In the linear case, the margin is the distance of the hyperplane to the nearest positive and negative examples.

$$u = \vec{w} \times \vec{x} - b \tag{3}$$

where w is the normal vector to the hyperplane and x is the input vector. The separating hyperplane is the plane u=0. The nearest points lie on the planes u = ±1.

The margin m is thus $$m = \frac{1}{\|w\|^2} \tag{4}$$

Maximizing margin is equivalent to minimize $$\frac{1}{2}\|w\|^2 \tag{5}$$

with the restriction: $$y_i \vec{w} \times \vec{x} - b \geq 1, \forall i \tag{6}$$

where x_i is the training example i, and y_i is the correct output of the SVM for training example i. The value y_i is +1 for the positive examples in a class and –1 for the negative examples.

Using a Lagrangian, this optimization problem can be converted into a dual form which is a quadratic programming (QP) optimization problem where the objective function Ψ is solely dependent on a set of Lagrange multipliers a_i .

$$\frac{1}{2}\sum_{N}^{i=1}\sum_{N}^{j=1} y_i y_j (x_i \times x_j)\alpha_i \alpha_j - \sum_{N}^{i=1}\alpha_i \tag{7}$$

(where N is the number of training examples), subject to the inequality constraints,

$$a_i \geq 0, \forall i \tag{8}$$

and one linear equality constraint, $$\sum_{N}^{j=1} y_i a_j = 0 \tag{9}$$

There is a one-to-one relationship between each Lagrange multiplier and each training example.

Once the Lagrange multipliers are determined, the normal vector $\vec{w}$ and the threshold b can be derived from the Lagrange multipliers:

$$\vec{w} = \sum_{N}^{j=1} y_i a_j \vec{x}_i, \quad b = \{\vec{w} \times \vec{x}_k - y_k, \quad for \quad a_k \neq 0 \} \tag{10}$$

Because $\vec{w}$ can be computed via equation (10) from the training data before use, the amount of computation required to evaluate a linear SVM is constant in the

number of non-zero support vectors. [1, 15]

Training a support vector machine requires the solution of a very large quadratic programming problem (referred further on as QP). It is not only time consuming but needs very large amount of memory and cannot be solved easily via standard QP methods. The quadratic form in (7) involves a matrix that has a number of elements equal to the square of the number of training examples.

Vapnik describes an optimization algorithm called “Chunking”. Chunking algorithm is based on the fact that the value of the quadratic form is the same if you remove the rows and columns of the matrix that corresponds to zero Lagrange multipliers. [18]

Therefore, the large QP problem can be broken down into a series of smaller QP problems, whose ultimate goal is to identify all of the non-zero Lagrange multipliers and discard all of the zero Lagrange multipliers.

Chunking algorithm scales the problem complexity between linear and cubic in the training set size.

Platt's Sequential Minimal Optimization (SMO) is a simple algorithm that can quickly solve the SVM QP problem without any extra matrix storage and without using numerical QP optimization steps at all. SMO decomposes the overall QP problem into QP sub-problems, using Osuna’s theorem to ensure convergence. The theorem proves that the large QP problem can be broken down into a series of smaller QP sub-problems. As long as at least one example that violates the Karush-Kuhn-Tucker conditions is added to the examples for the previous sub-problem, each step will reduce the overall objective function and maintain a feasible point that obeys all of the constraints. [13].

Unlike the previous methods, SMO chooses to solve the smallest possible optimization problem at every step. For the standard SVM QP problem, the smallest possible optimization problem involves two Lagrange multipliers, because the Lagrange multipliers must obey a linear equality constraint. At every step, SMO chooses two Lagrange multipliers to jointly optimize, finds the optimal values for these multipliers, and updates the SVM to reflect the new optimal values.[8]

The advantage of SMO lies in the fact that solving for two Lagrange multipliers can be done analytically. Thus, numerical QP optimization is avoided entirely. Even though more optimization sub-problems are solved in the course of the algorithm, each sub-problem is so fast that the overall QP problem is solved quickly. Also SMO is linear in the training set size.

In addition, SMO requires no extra matrix storage at all so the amount of necessary memory is greatly reduced. Furthermore, since no matrix algorithms are used, SMO is less susceptible to numerical precision problems.[9,15].

1.5 Evaluation

Since our approach is based on document classification we will be using the two

main information retrieval indicators - ***precision*** and ***recall***.

Precision is a measure of how many relevant document are retrieved from the total number of documents retrieved (relevant or not) and recall is the number of relevant document are retrieved from the total number of relevant documents. [1]

$$Recall = \frac{tp}{tp + fn} \tag{11}$$

$$Precision = \frac{tp}{tp + fp} \tag{12}$$

where:

tp = true positives - documents correctly classified as included in one document class

fn = false negatives - documents belonging to that specific class that have not been found

fp = false positives – incorrect documents classified as belonging to that class

1.6 Algorithm

The main differences between organic farming methods and classic agriculture methods are in crop fertilization and pests control but the methods to plant, grow and gather crops are common. Therefore, it was necessary to build a fine grained classification model.

Moreover often same document contains both organic and classic agriculture best practices or organic agriculture tips expressed negating classic agriculture methods. From this perspective we prepared two sets of classification models: one at phrase level and the second at document level.

Our model has four main steps: corpus preprocessing, training, testing and evaluation.

Training and testing corpus preprocessing

Filter stopwords (less than 3 characters); Extract only relevant concepts from both training and testing using the AGROVOC ontology. AGROVOC contains around 535.000 concepts from witch 40.000 English concepts. For extracting such concepts Porter stemming algorithm is also used.[16]

The general training model can be described in the following steps:

Extract features from training corpus for each class; Calculate Tf*Idf measure for each feature; Build Word Vector representation of each; document using feature, Tf*Idf measure pairs; Build the training model.

Document testing model follows these steps:

Extract features from each document of testing corpus; Calculate Tf*Idf measure for each feature; Build Word Vector representation of each phrase using feature, Tf*Idf measure pairs; Evaluate each document against the training model and assign a class;

Phrase testing model main steps:

Divide each document into phrases using Chaos; Extract features from each phrase of testing corpus; Calculate Tf*Idf measure for each feature; Build Word Vector

representation of each phrase using feature, Tf*Idf measure pairs; Evaluate each phrase against the training model and assign a class; Calculate each document score for both organic and no-organic number of phrases; Assign the class to each document based on the previously calculated score.

Evaluate results:

Calculate precision and recall for both methods: document based classification and phrase based classification.

2 Experiment Configuration

Following several tests we empirically determined that for our data the best results were obtained using a linear SVM. Moreover we found out that the following SMO configuration gave the best results:

SMO Parameters: C 1.0; L 0.0002; P 1.0E-12; N 0; V -1; W 1; E 1.0; K PolyKernel; **Kernel parameters:** C 250007

Where:

C The complexity constant C.

E The exponent for the polynomial kernel.

G Gamma for the RBF kernel.

N <0|1|2> Whether to 0=normalize/ 1=standardize /2=neither.

F Feature-space normalization (only for non-linear polynomial kernels).

O Use lower-order terms (only for non-linear polynomial kernels).

R Use the RBF kernel. (default poly)

A Sets the size of the kernel cache. Should be a prime number.

L Sets the tolerance parameter. (default 1.0e-3)

P Sets the epsilon for round-off error.

M Fit logistic models to SVM outputs.

V Number of folds for cross-validation used to generate data for logistic models.

W Random number seed for cross-validation.

3 Experiment Results

In this section we analyze results of several experiments in agriculture and organic agriculture text classification. Firstly we will discuss the results of the tests on the manually selected corpus then the results on the web extracted documents. The aim was to compare results obtained from the classification of the first well focused corpus with the second real life corpus.

3.1 First corpus experiments

As mentioned in sections 4.2 and 4.3 the training corpus was composed by 576 documents (5061 phrases) in agriculture class and 96 documents (1322 phrases) for the organic class. First testing corpus was made by 168 documents (1602 phrases). Organic and no-organic documents identification results are shown in *Table 1*.

As shown in *Table 1* we identified correctly all no-organic documents and a good number of organic documents - 19 from a total of 24: in percent a little over 79%.

In *Figure 1* are shown the two main indicators of document classification: precision and recall. We got impressive results on no-organic classification (100% recall and 97% precision) and very good ones on organic document recognition (79.1% recall and 100% precision).

Precision and recall indicates that on this first experiment we have the ability to correctly classify nearly 80% of organic documents and that the probability to find firstly the most relevant topics first is 100%. We also identify correctly all no-organic documents with a precision of 97%.

ORGANIC DOCUMENT IDENTIFICATION	
Organic identified as noorganic	5
Noorganic identified as organic	0
Organic correctly identified	19
Noorganic correctly identified	144
TOTAL ORGANIC TESTED	24
TOTAL NOORGANIC TESTED	144
TOTAL TOPICS TESTED	168

Table 1: First experiment set results

ORGANIC DOCUMENT IDENTIFICATION	
Organic identified as noorganic	0
Noorganic identified as organic	17
Organic correctly identified	12
Noorganic correctly identified	55
TOTAL ORGANIC TESTED	12
TOTAL NOORGANIC TESTED	72
TOTAL TOPICS TESTED	84

Table 2: Second experiment set results

From an economic point of view is a very good result as the system is capable of finding relevant documents on two very close or even overlapped document classes. We have to mention here that it is remarkable that the same classification algorithm ran over the corpus composed by document phrases gives exactly the same results obtained on full documents classification.

The phrase based classification score was computed calculating the number of correctly identified phrases for each document. If the document scores over 50%

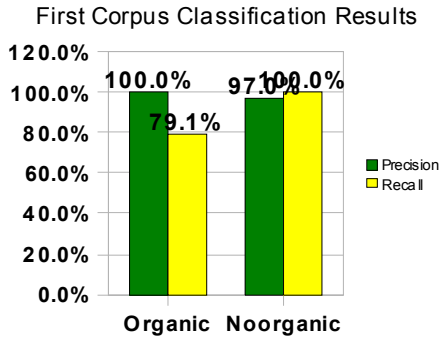

Figure 3: First Corpus Precision and Recall

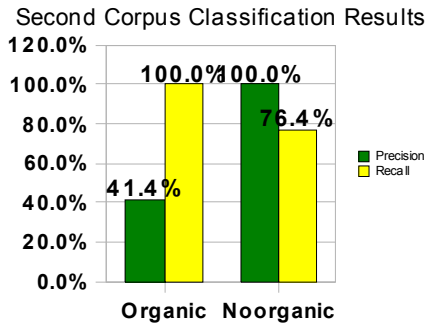

Figure 2: Second Corpus Precision and Recall

correctly classified phrases it is considered correctly classified. This shows that the algorithm we implemented is very robust and stable over the same corpus even if we work at document level or phrase level.

3.2 Second corpus experiments

For the second set of experiments training corpus is the same as for the first set of experiments while testing corpus contains 84 documents (1593 phrases) from witch 72 no-organic documents and 12 organic documents to maintain the exact proportion as for the first experimental setup.

Web documents are longer than the manually processed ones – contain nearly two times as much phrases. Organic and no-organic documents identification results are shown in *Table 2*. As shown in *Table 2* we identified correctly all organic documents and a good number of no-organic documents - 55 from a total of 72. In percent is over 76%.

In *Figure 2* it is shown the two main indicators of document classification: precision and recall. We obtained promising results on organic classification (100% recall and 41.4% precision).

It was expected a drop in precision since the documents are directly gathered from the web with virtually no human in depth manual inspection. We also got good results on no-organic document recognition (76.4% recall and 100% precision).

This results could be improved with a more complete training corpus. Often the documents directly gathered from the web contain more concepts than training corpus so it is virtually impossible to get a better score.

From an economic point of view is a promising result as the system is capable of finding relevant documents on both classes. We have to mention here that as in the first experiment setup the results obtained by classifying individual document phrases and then aggregating results are exactly the same as the results obtained on full documents classification.

4 Conclusions and Future Work

In this paper we presented our study on agriculture and organic agriculture document classification with ontology support.

We introduced some general notions and the current state of document classification with the support of Ontologies. Then we detailed our model including document representation, corpus description, theoretical background, evaluation, and proposed algorithm.

Furthermore we widely described our experimental setup and results. As shown in section 6 for both experiment sets we obtained really good results. The system is capable of classifying two very close classes of documents on two different testing corpora. As mentioned in section 4.3 the first corpus is a fairly specific corpus but wasn't built by domain experts and the second corpus is entirely gathered from the web with no human intervention. We expect that with a better and more complete training corpus organized by domain experts results could be further improved.

For these first experiments in the agriculture domain we used only little support from the FAO Ontology: mainly as a domain thesaurus. The system could be further improved by using the full capabilities of FAO ontology as heuristics.

Another future research direction could include multilingual information retrieval based on AOS as a multilingual agriculture ontology and ontology enrichment. It may be possible to enrich AGROVOC ontology with the specific concepts of organic agriculture extracted with the aid of text mining engine from training and testing corpus. Also phrase construction and semantic orientation could be further investigated.

Furthermore, on top of this models a portal should be build to integrate and share world wide organic agricultural information and especially best practices. The portal should bring together local experiences, standards and various document summaries and therefore to create a common knowledge base. [9]

REFERENCES

[1] **Basili, R., Moschitti, A. (2005**), ***Automatic Text Categorization: From Information Retrieval to Support Vector Learning***, Aracne Editrice, Informatica, ISBN: 88-548-0292-1;

[2]**Bodea, C. (2007)** ***Artificial Intelligence Techniques Applied to the Evaluation of the Research and Technology Development Projects and Programmes***, Journal of Economic Computation and Economic Cybernetics Studies and Research, 2007, Issue 3-4, ASE Publishing House, Bucharest;

[3] **Carroll, J. J., Dickinson, I., Dollin, C., Reynolds, D., Seaborne, A., Wilkinson, K. (2004),** ***Jena: Implementing the Semantic Web Recommendations.*** In Proceedings of the 13th international conference on World Wide Web - Alternate Track Papers \& Posters, WWW 2004, New York, USA, May 17-20, 2004, 74-83;

[4]**Chen, D., Müller, H-M., Sternberg, P. W. (2006),** ***Automatic Document Classification of Biological Literature.*** BMC Bioinformatics.; 7: 370;

[5] **Joachims, T. (2002),** ***Optimizing Search Engines Using Clickthrough Data***, Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD), ACM;

[6] **Joachims, T. (1998),** ***Text Categorization with Support Vector Machines: Learning with Many Relevant Features.*** Proceedings of the 10th European Conference on Machine Learning 1998, 137-142;

[7]**Kwang Mong Sim (2004),** ***Toward an Ontology-enhanced Information Filtering Agent.*** 33 (1) 95 – 100, ACM New York, NY, USA;

[8] **Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., Murthy, K. R. K. (1999),** ***Improvements to Platt's SMO Algorithm for SVM Classifier Design***. Technical Report CD-99-14. Control Division, Dept of Mechanical and Production Engineering, National University of Singapore;

[9] **Lungu, I.. Velicanu, M., Bara, A., Diaconita, V., Botha I. (2009),** ***Portal Based System Integration – Foundation for Decision Support***. Journal of Economic Computation and Economic Cybernetics Studies and Research, Issue 1, ASE Publishing House, Bucharest;

[10] **Lungu, I., Manole, S. (2007).,*****Estimation of Nonparametric Regression Function.*** Journal of Economic Computation and Economic Cybernetics Studies and Research, Issue 1-2, ASE Publishing House, Bucharest;

[11] **Muller, H. M., Kenny, E. E., Sternberg, P. W. (2004),** ***Textpresso: An Ontology-based Information Retrieval and Extraction System for Biological Literature,*** PLoS Biol. 2004 Nov;2(11):e309. Epub Sep 21;

[12] **Nagarajan, M., Sheth A., Aguilera, M., Keeton, K., Merchant, A., Uysal, M. (2007),** ***Altering Document Term Vectors for Classification - Ontologies as Expectations of Co-occurrence.*** In Proceedings of the 16th International Conference on World Wide Web, Banff, Alberta, Canada; POSTER SESSION: Semantic web. 1225 – 1226;

[13] **Osuna, E., Freund, R., Girosi, F. (1997),*****Training Support Vector Machines: An Application to Face Detection***, Proc. Computer Vision and Pattern Recognition '97, 130-136;

[14] **Pazienza, M. T., Stellato, A.,Tudorache, A. (2008),** ***Flames, Risky Discussions, No Flames Recognition in Forums.*** EMOT. The Book!, To be published by Springer-Verlag;

[15]**Platt, J. (1998),** ***Fast Training of Support Vector Machines using Sequential Minimal Optimization. Advances in Kernel Methods - Support Vector Learning***, B. Scholkopf, C. Burges, and A. Smola, eds., MIT Press;

[16] **Porter, M.(1980)** , ***An Algorithm for Suffix Stripping. Program***, 14:130-137;

[17] **Subirats, I., Onyancha, I., Salokhe, G., Kaloyanova, S., Anibaldi, S., Keizer, J. (2008),** ***Towards an Architecture for Open Archive Networks in Agricultural Sciences and Technology.*** Online Information Review, Vol: 32, Issue 4, Pages 478 – 487;

[18] **Vapnik, V. (1982),** ***Estimation of Dependences Based on Empirical Data***, Springer-Verlag;

[19] **Witten, I. H., Frank, E. (2005),** ***Data Mining: Practical Machine Learning Tools and Techniques*** (Second Edition). Morgan Kaufmann.