An Approach to Learning Objects Classification using an Auto-Associative Neural Network

Clara Pérez-Molina and Miguel Rodríguez-Artacho

UNED ETSI Industriales / UNED ETSI Informática
Juan del Rosal, 16 28040 Madrid (Spain)
clarapm@ieec.uned.es , miguel@lsi.uned.es

Abstract. In this paper we present an unsupervised neural network-based approach to multiword extraction and semantic classification of learning material with potential application to automatic metadata generation and ontological classification of learning objects. The described experience shows the behavior of a neural network-based system to extract semantic units in a given learning material. Semantic units contribute to a better indexing and therefore characterization of the dataset making easy the access to the knowledge contained in the course. Extraction of these semantic units identifies the content from an instructional perspective and can then be applied to instructional classification according to a given ontology.

1 Introduction

Meaningful words found together in a variety of contexts can express very accurately the semantic of a given text. These items, also known as multiword expressions or complex concepts, consist of two or more words that correspond to some conventional way of referring topics or concepts [1]. When we search information about a knowledge item, the subject is usually described by a set of words; this is what we call semantic unit. When applying semantic units' extraction to the educational content arena, we find that the help of existing ontologies and educational classifications can allow a framework to automatically scaffold chunks of content according to these conceptualizations.

In this context, classification of learning material is generally done providing a precise set of words that accurately describe a given resource. As we know, current authoring process of learning content is difficult and costly as it combines search and retrieval of the appropriate Learning Objects (LO) to compose larger units of content [9]. Current standards as IMS or SCORM label with a rich metadata profile the new resource which is manually elaborated and this result in a trade-off between low granularity resources, very reusable, but costly and major pieces not as reusable due to high context dependant information.

Thus approaches to automatic semantic labelling of content are becoming an important objective in order to achieve automatic metadata generation as current model of manually labelled material has been frequently reported to be difficult to maintain [7,8]. Some approaches have been developed by means of clustering of

© A. Gelbukh, Á. Kuri (Eds.)
Advances in Artificial Intelligence and Applications
Research in Computer Science 32, 2007, pp. 395–405

Received 17/06/07
Accepted 31/08/07
Final version 30/09/07

learning material using formal concept analysis [10], which eventually can capture the knowledge representation of the learning content as a lattice, similar to a real taxonomy.

In this paper we propose to apply a parallel approach using an associative neural network self trained with non supervised learning to the classification of a given educational material. Our aim is to explore non supervised training of our neural network with the intention to obtain a set of semantic units that could describe any educational material and classify it according to a given instructional classification, as for instance the one developed in [9]. The rest of the paper is structured as follows: section 2 explains our neural network design and the semantic units' extraction. Section 3 applies the model to a given course, extracting the neural network results. Finally section 4 shows the conclusions.

2 Classification System Design

At first glance, the simplest method for finding out semantic units is counting, therefore, in many cases researchers rely on frequency methods. However, neither just selecting the most frequent bigrams (sequences of two adjacent words), nor considering sets of part of speech tag patterns (Adjective-Noun, Noun-Noun, Adjective-Adjective-Noun,...) are good solutions, due to the noisy words that sometimes appear between connected words [2]. On the other hand, syntactic methods need a previous tagged phase, which sometimes is not possible to carry out and, moreover, it may introduce errors to the rest of the process [3].

The nature of management information problem in text documents has provided an important area of application for some kinds of neural networks [4]. Unsupervised neural networks as Kohonen networks [5], Hopfield networks or hybrid systems [6] have been implemented.

Our intention is that the developed system generates behavior pattern matrixes from term sequences in documents that are used to train an auto-associative neural network. After the training stage, the network owns information that involves every atomic concept appeared in the text. The analysis of neuron connections provides a method for detecting semantic units inspecting the heaviest weights in the network, since they are consequence of most repeated behavior patterns. In this way, the model allows to identify sequences of words that represent semantic units in the dataset processed.

2.1 Generation of Patterns

Each neuron in the network represents one atomic concept, that is, a concept that can be represented by just one content word. In order to carry out this correspondence between neurons and atomic concepts, a previous analysis stage is required. Such analysis stage has two objectives. The first one is to identify which words can represent atomic concepts. The second one is to ensure that whenever a particular concept appears it is always represented by the same neuron. In order to do that, a

stemming process is applied to the content words found in the text. (See Figure 1, where only first 100 of 487 files processed is shown)

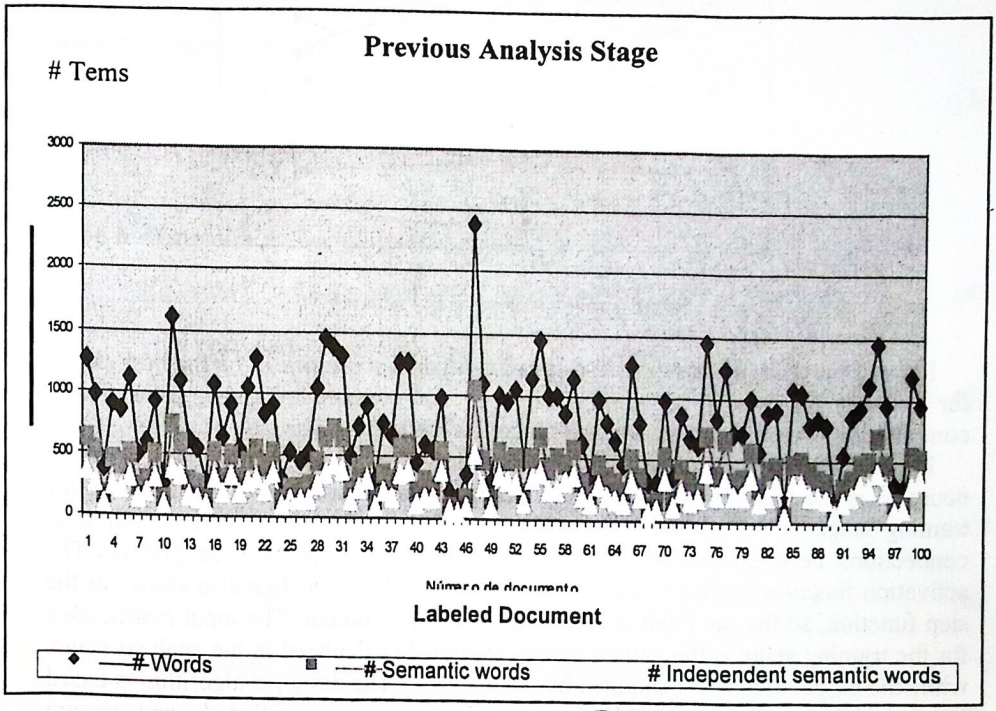

Fig. 1. Previous Analysis Stage: Independent semantic words identification.

The pattern generation phase starts once atomic concepts are identified in the text. In, order to generate the behavior pattern matrix associated to the text, word environments are considered through a scrolling window that is moving over the text. The window size is related to the maximum number of atomic concepts that constitute semantic units detected by the system; however this parameter can be easily modified.

After concluding this stage, we obtain a behavior matrix M that represents the entire text, whose dimensions are defined by SxQ. S is the number of atomic concepts found in the text and Q is the number of generated patterns. The set of patterns are grouped according to their generation order to shape the column vectors of M, taking into account that each pattern is originated by one scrolling window position.

2.2 The Neural Network Algorithm

The neural network architecture selected for the system is a fully connected neural net. Hence, each neuron is connected to the rest of them, including itself.

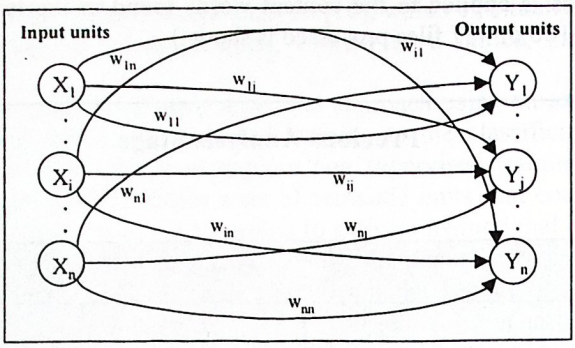

Fig. 2 Autoassociative neural network.

For our purpose, an unsupervised learning algorithm seemed to be the best choice for training the neural network, since it is not possible to know, prior to the computation, which is the appropriate output for the net.

Initially the weight matrix, that reveals the strength of connections between neurons, is the identity matrix, so each neuron is just associated to itself. When the training stage is completed, the elements of the weight matrix corresponding to connections between neurons that make up semantic units will be greater. The activation function applied to the network is the hard limit function also known as the step function, so the net input is converted to a binary output. The input matrix used for the training stage is the pattern matrix previously calculated in the analysis stage, while the output matrix is computed by the network. The designed algorithm is based on the Hebbian learning [11]. Two learning parameters called *lr* and *dr* are introduced. The first one represent the learning rate, the second one the decay rate. These parameters can be adjusted to achieve the best configuration according to the dataset. The developed algorithm for an autoassociate neural network of S neurons can be described by the following steps:

Step 0: Firstly, it is necessary to calculate the dimension of the network. Considering M as the representative matrix of the text, there are as many neurons as rows, S.

Step 1: The weight matrix of the net $W=\{w_{ij}\}$ is initially the identity matrix.

$$W = W_0 = I \qquad \begin{matrix} w_{ij} = 0 & \text{if } i \neq j \\ w_{ij} = 0 & \text{if } i = j \end{matrix} \tag{1}$$

Step 2: A bias term, whose value is controlled through a parameter α, is introduced. The bias term B is expressed by a vector with as many components as neurons constitute the network. Initially,

$$B = B_0 = -\alpha \, \mathrm{u} \qquad b_i = -\alpha \tag{2}$$

Step 3: For T times, do steps 4-7: ($\tau = 1,...,T$)

Step 4: For each of the Q training vector P(q) do steps 5-7: ($q = 1,....,Q$)

Step 5: Compute net input,

$$Y\_in(q) = W * P(q) + B \qquad y\_in_i(q) = \sum_{j=1}^{S} w_{ij} * p_j(q) + b_i \tag{3}$$

Step 6: Determine activation (output signal),

$$A(q) = f(Y\_in(q)) = hard\lim(Y\_in(q)) = hard\lim(W * P(q) + B) \tag{4}$$

For each unit, the activation is

$$a_i(q) = hard\lim\left(\sum_{j=1}^{S} w_{ij} * p_j + b_i\right) \tag{5}$$

Step 7: Update weights,

$$W^{\tau} = W^{\tau-1} + dW^{\tau} \qquad w_{ij}^{\tau} = w_{ij}^{\tau-1} + dw_{ij}^{\tau} \tag{6}$$

The weight matrix W of the net changes by an amount dW given by,

$$dW^{\tau} = lr * A(q) * P^{T}(q) - dr * W^{\tau-1} \qquad dw_{ij}^{\tau} = lr * a_i(q) * p_j(q) - dr * w_{ij}^{\tau-1} \tag{7}$$

Training stage is finished when the number of iterations has reached a determined value T, this parameter needs to be adjusted for each dataset. Information about atomic concept connections is then contained in the weight matrix of the network, so that it is necessary to develop an analysis mechanism to find them out.

2.3 Semantic Units' Extraction

The analysis method applies a filter process to the weight matrix, so that every element not exceeding a threshold μ is rejected. Elements belonging to the main diagonal are also removed in the filter process, since they do not involve information (according to our exposition, elements of the main diagonal represent the relationship between one atomic concept and itself) (Figures 3 and 4).

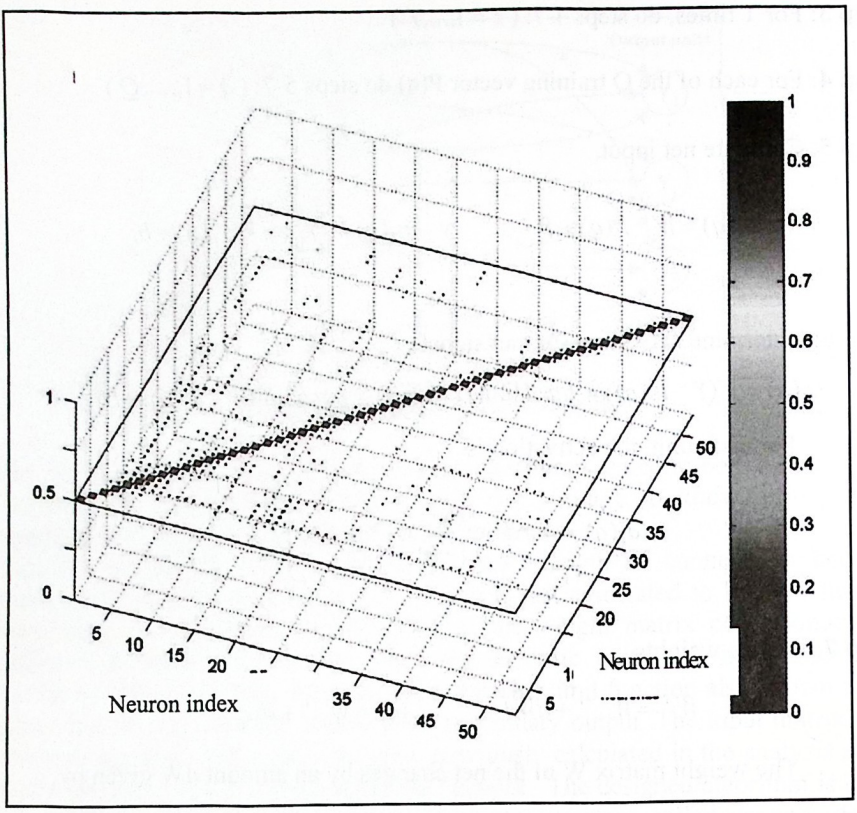

Fig. 3. The figure shows the visualization of the weight matrix. In weight matrix representation, x and y axes represent neuron indexes and z axe the corresponding weight for each connection. This example is extracted of a network of 53 neurons.

When the filter process to the weight matrix is concluded, the extraction of semantic units' stage begins

The extraction stage is carried out in different phases; each one is in charge of the extraction of semantic units of a particular number of atomic words. The developed system is able to extract semantic units made of two, three, four and five atomic concepts. First of all, the system search for semantic units made of two terms, and after the number of terms is increased one by one. The search algorithm takes into account that the weight matrix is not symmetric. Due to learning algorithm characteristics used for training the neural network, the weight on connection from neuron i to neuron j is different from the weight on connection from j to i. Therefore, the search algorithm imposes the criterion of reciprocity. It establishes that if neuron i and j make up a couple it is necessary that both weights w_{ij} and w_{ji} exceed the thresholdμ. When couples are extracted, the search algorithm concerns with complex concepts of three terms. The algorithm assumes that every concept belonging to a set of three elements makes up couple with the rest of them. Finally, the extraction of semantic units of four and five terms is carried out in the same way.

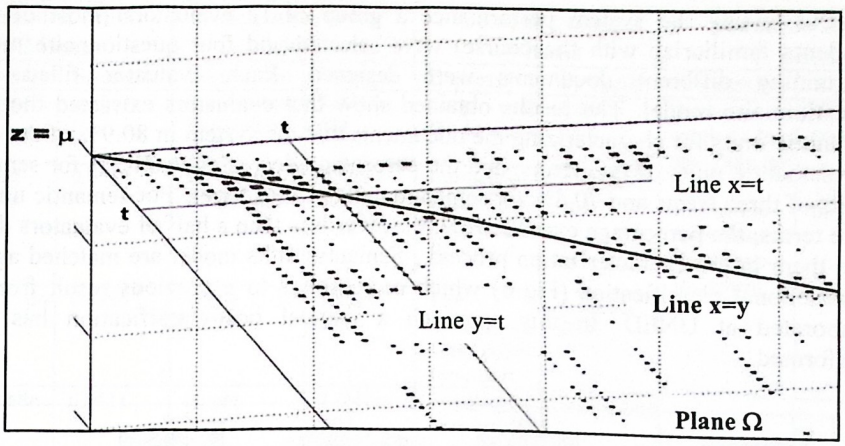

Fig. 4. The figure shows the visualization of the filter process as the projection of a three dimensional space representation of the weight matrix over the plane z=μ; x and y axes represent neuron indexes.

3 An Approach to Classify Learning Material: Experimental Results

The data used in our experiments consists of a collection of 487 documents concerning to different Java aspects of a generic Java course[1] arranged as a freely available web site.

Taking into account the formatting stage applied to the files, the semantic units extracted are made up of terms that do not appear in the text in the same form (or at least the majority of them). The identification process of the semantic units associated to each document of the dataset, requires an additional process by which the system applies an ordering method and assigns a linguistic expression to each group.

	# words	# Filtered words	# indep. terms	# dependent terms	# patters	Document Representative matrix dimension
Arithmetic mean	153	69	48	36	79	48 x 79
Median	136	62	46	29	71	46 x 71
Standard Deviation	67	34	16	23	36	16 x 36
Variance	4535	1134	282	423	1312	282 x 1312

Fig. 5. Semantic units' results classified by words, terms and patterns

[1] In this experiment our source is "Introducción a Java" course. We thank DISPEIL project, UNED and Universidad Carlos III de Madrid.

For testing the system performance a group of 12 evaluators (most of them students familiarize with the course) were selected and four questionnaire models containing different documents were designed. Each evaluator filled in a questionnaire model. The results obtained show that evaluators extracted the same semantic units for characterizing the documents that the system in 80.9% of the cases for semantic units of two terms, and the percentage decreases to 71.2% for semantic units of three terms and 70.4% for semantic units of four terms. For semantic units of five terms, the percentage exceeds 60%, however less than a half of evaluators do not use them in the characterization process. Semantic units model are matched against instructional classification (Fig 6) which corresponds to a previous result from [9] elaborated at UNED. In this approach a manual post-classification has been performed.

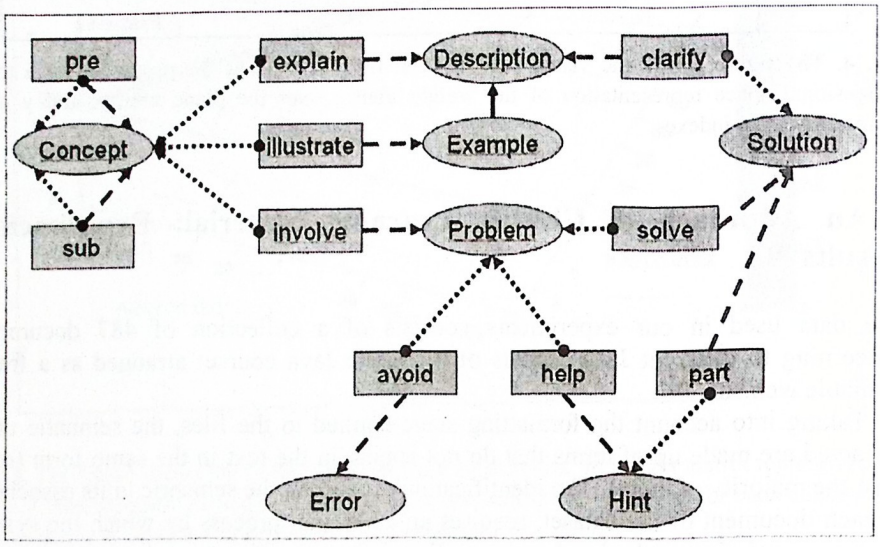

Fig. 6. The existing instructional conceptualization

Multi-words consistent of course related terms are searched in a Java vocabulary and then manually classified in the ontology. The main interest has not been so far to perform at this stage a full automatic process of classification but to explore how accurately semantic units' extraction can be mapped against an ad-hoc classification [14].

For instance, let us consider an example obtained by processing through the semantic network a given file. Processing results are shown in Table 1 where semantic units of size 1 to 5 are shown. Multi-words consistent of java related topics (as reserved java words like `for`, and `while`) are classified as *example of* the given reserved word, thus file is associated to *EXAMPLE* class that *illustrate* concepts thus producing a link semantically relevant between the plain resource and the instructional ontology.

Semantic units: 2 items		Semantic units: 3 items		Semantic units: 4 items		Semantic units: 5 items	
Expression	Strength	Expression	Strength	Expression	Strength	Expression	Strength
Statement **while**	0.0547	Initialization finalization increment	5.0605e-003	Loop **while** for **do-while**	8.648e-0.04	Sentence **while** loop **do-while for**	3.4877e-005
Loop **for**	0.0539	Sentence repeat loop	4.4703e-003	Instruction Initialization Finalization Increment	6.4251e-004	Loop instruction Initialization Finalization Increment	1.4165e-005
Loop **do-while**	0.0495	**While** for do-while	3.4613e-003	**For** Initialization Finalization increment	4.3869e-004		
Loop **while**	0.0411	Iterate Increment loop	2.3686e-0.03	Finalization sentence loop expression	3.147e-004		
do-while	0.0372	**While** expression sentence	1.9725e-003				
Repeat	0.0320	**For** expression sentence	1.7148e-003				
Initialization sentence	0.0285	Do-while expression sentence	1.623e-003				
Termination Sentence	0.0263	Initialization sentence loop	0.9364e-003				
Initialize loop	0.0192	End Sentence loop	0.8295e-003				
Finish loop	0.0187						
Increment expression	0.0174						
Increment counter	0.0169						
Execute sentence	0.0158						
Construction loop	0.0136						
Evaluate expression	0.0114						

Table 1. Semantic units' analysis from file `text_ch5s01s01.html`

4 Conclusions

The described model provides a new approach based on unsupervised neural networks for detecting semantic units constituted by up to 5 atomic concepts that can be applied to different datasets of educational material: full-length document collections or

single documents. The characterization obtained can be used for content identification purposes including information retrieval systems, thesaurus generation or issues related to natural language generation.

The regularities of human language studied allow characterizing text documents in a better way, since this approach stresses the importance of multiword concepts instead of atomic words, and there is usually an element of meaning added to the combination of words [12] that make up a multiword. On the other hand, it holds similarities with the way human mind is supposed to work. In addition to a better semantic characterization, the study of these regularities to extract multiwords produce important savings in the number of required indexing terms to represent documents, what is traduced in savings in the amount of software memory or electronic circuitry devoted to this task.

The conclusions obtained in this work suggest that the proposed algorithm could be an interesting tool in automatic semantic units identification processes. The main inconvenience of other approaches that rely on sets of part of speech tag patterns is that noisy words that can appear between connected words avoid the multiword extraction [13], however the neural network design proposed is not subordinated to such limitation since it is not affected by words appeared between two connected. As the evaluation process shows, the system is particularly effective for detecting complex concepts made up of not only two terms but also, made up of three, four and five terms, what is particularly interesting for a better characterization of documents. The inconvenience is that the term order inside a group is lost. However a statistical process is carried out in order to solve the problem [16]. The method accuracy is rather high, the system is able to detect neuron associations due to patterns that just appear a few times in the dataset. The criterion of reciprocity used avoids that false couples provided by noise pattern (patterns resulting from chance events) could be make up, even if the selected threshold is too low.

Future work will also trend to perform more experiments using richer input like open courseware of UNED University. Also will use experience from other similar works on automatic metadata extraction and classification, as mentioned in [7] to explore non-labeled open web courses.

References

1. Manning, C., Schütze, H.: Foundations of Statistical Natural Language Processing. The MIT Press. Cambridge, Massachusetts. London, England. (1999) pp. 151-189.
2. Justeson, J., Katz, S.: Technical terminology: some linguistic properties and an algorithm for identification in text. Natural Language Engineering. (1995) pp. 9-27.
3. Chen, H., Lynch, K.: Automatic Construction of Networks of Concepts Characterizing Document Databases. IEEE Transactions on Systems, Man, and Cybernetics, 22. (1992) pp. 885-902.
4. Ruiz, M., Srinivasan, P.: Automatic Text Categorization Using Neural Networks. Advances in Classification Research, 8: Proceedings of the 8th ASIS SIG/CR. Ed. Efthimis Efthimiadis. Information Today, Medford, New Jersey. (1998) pp. 59-72.
5. Kohonen, T.: Self-organization and associative memory. Springer Series in Information Sciences, Vol. 8. Springer-Verlag, New York. (1984)

6. Lin, C., Chen, H.: An automatic Indexing and Neural Network Approach to Concept Retrieval and Classification of Multilingual (Chinese-English) Documents. IEEE Transactions on Systems, Man and Cybernetics: Cybernetics, 26(1) (1996) pp. 75-88.
7. C. D. Kloos and M. Rodríguez-Artacho (2004) "Semantic Web aided Authoring of Educational Material" in proceedings of the SW-EL 06 conference
8. J. Qin & N Hernández (2004) "ontological representation of learning objects: building interoperable vocabulary and structures" in proceedings of WWW2004 conference. ACM press
9. M. Rodríguez-Artacho and F. Verdejo (2004) "Modelling Educational Content: The cognitive approach of PALO language" In journal of Educational technology and society Vol 7 (3), 2004 124-137.
10. Mohan, P. Brooks, C (2003) "Learning objects on the semantic Web". proceedings of the IEEE conference on advanced learning technologies ISBN: 0-7695-1967-9
11. MacLeod, K., Robertson, W.: A Neural Algorithm for Document Clustering. Information Processing and Management, 27 (3). (1991) pp. 337-346.
12. Firth, J. R.: A synopsis of linguistic theory 1930-1955. Studies in Linguistic Analysis. Oxford: Philological Society. (1957) pp. 1-32.
13. Smadja, F.: Retrieving collocations from text: Xtract. Computational Linguistics. (1993) pp. 143-177.
14. Caropreso, M. F., Matwin, S., Sebastiani, F. (2000). Statistical phrases in automated text categorization. Technical Report, Instituto di Elaborazione dell'Informazione, Consiglio Nazionale delle Recerche.
15. Lin, X., Soergel, D., Marchionini, G.: A Self-Organizing Semantic Map for Information Retrieval. Proceedings of the 14th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval. (1991) pp. 262-269.
16. Perez C. "Intelligent system design using artificial neural networks for automatic semantic extraction in medical literature. Second International Workshop on Intelligent System Design and Applications ISDA 2002. Atlanta, USA, August 2002.