Computer and Information Science; Vol. 11, No. 1; 2018
ISSN 1913-8989 E-ISSN 1913-8997
Published by Canadian Center of Science and Education

A Class Validation Proposal of a Pedagogic Domain Ontology based on Clustering Analysis

Yuridiana Alemán[1], María J. Somodevilla[1] & Darnes Vilariño[1]

[1] Faculty of Computer Science, Benemérita Universidad Autónoma de Puebla, Puebla, México

Correspondence: María J. Somodevilla, Faculty of Computer Science, Benemérita Universidad Autónoma de Puebla, Puebla, México. Tel: 52-222-229-5500. Ext 7228. E-mail: mariajsomodevilla@gmail.com

Received: December 17, 2017 Accepted: December 27, 2017 Online Published: January 16, 2018

doi:10.5539/cis.v11n1p65 URL: http://dx.doi.org/10.5539/cis.v11n1p65

Abstract

The knowledge bases of the Web are fundamentally organized in ontologies in order to answer queries based on semantics. The ontologies learning process comprises three fundamental steps: creation of classes and relationships, population and evaluation. In this paper the focus includes the classes creation, by introducing a class validation proposal using clustering analysis. As case of study was selected a pedagogical domain, where a corpus was semi-automatically built, from articles written in Spanish published in Social Sciences. Moreover, a dictionary containing classes, concepts and synonyms was included in the experiments. Clustering analysis allowed to verify the concepts that the experts considered as the most important for the domain. For the case of study selected, the cluster analysis step reports clusters with the same instances that the clusters defined by the experts.

Keywords: NLP, corpus, dictionary, ontology, class validation, pedagogical domain, learning styles, intelligences types learning strategies

1. Introduction

Nowadays, the available information has increased exponentially, therefore it is necessary to propose novel techniques or processing that information and use them with different scientific objectives. The ontologies can be used for purposes such as structure knowledge in taxonomies, vocabulary management, natural language processing applications, searches, recommendation systems, e-learning among others (El-Ansari et al., 2016).

The ontology learning process integrates the class detection, creation, population and evaluation. This process is applied in different researches such as Fu et al. (2008), García et al. (2010) and Ochoa et al. (2011). The ontology learning process needs resources compatible with the research. This paper is focused on the initial two steps of ontology learning process: the corpus analysis and principal classes detection. A dictionary containing principal concepts and a corpus formed with pedagogical papers in Spanish language were built in the process. Pedagogical domain is extensive, thus the research is focused on the creation of tools that support the classes of the instructors in the classroom. Three topics, according tha opinion of experts, were widely researched: learning styles, intelligences types and learning strategies in order to be considered as the main classes of the ontology.

The article is organized in eight sections described as follows. Section 2 introduces the ontology learning problem and the ontology formalization process. Section 3 describes the topics addressed in the corpus, section 4 discusses the clustering methods and the available evaluation metrics being used for experiments. Section 5 presents related work about the pedagogical domain mainly and some other domains; besides works done on ontology learning process are also discussed. Section 6 present the proposed methodology and section 7 shows the analysis of results. Finally, section 8 presents conclusions and future work of the research.

2. Ontology Learning

Ontology is a philosophical discipline which can be described as the science of existence or the study of being. In modern computer science parlance, an ontology is a formal and explicit specification of a shared conceptualization of an interest domain. Their classes, relationships, constraints and axioms define a common vocabulary to share knowledge (Guarino et al., 1999). This is such as a data base, and is defined with an initial corpus where the principal components or keywords are extracted. Afterward, the relationships between keywords are inferred and a graph structure is created (the keywords are the nodes and the relationships are the edges).

Ontologies can be used for purposes such as structure knowledge in taxonomies, vocabulary manage, natural language processing applications (El-Ansari et al., 2016), searches (Celjuska & Vargas-Vera, 2004), recommendation systems (Dai & Li, 2010), and e-learning. Ontologies can model interaction systems between users and their environment, since to its property to manage complex knowledge in reusable formal representations. Formally, ontology can be defined such as (Faria et al., 2014):

$$O = (C, H, I, R, P, A)$$

Where:

- C is the set of entities of the ontology
- H is the set of taxonomic relationships between concepts
- I is the set of instance relationships related to the C
- R is the set of non-taxonomic ontology relationships
- P is the set properties of ontology entities
- A is the set of axioms, rules that allow checking the consistency of an ontology and infer new knowledge through some inference mechanism

The process called ontological learning is carried out to generate knowledge and includes the following tasks (Cimiano, 2006):

1. Acquisition of relevant terminology
2. Identification of synonyms terms
3. Formation of concepts
4. Hierarchical organization of concepts (concept hierarchy)
5. Learning relations, properties or attributes, along with the appropriate domain and rank
6. Hierarchical organization of relations (relation hierarchy)
7. Instantiation of axiom schemata
8. Definition of arbitrary axioms

Ontology population is usually accomplished through out three stages: Identification of *candidate instances, classifier construction and instances classification.* The input is an ontology and a corpus where the candidate instances are identified. Using a classifier, the instances are labeled with a class and finally, the output is the ontology populated (Faria et al., 2014).

The mainly problem faced in ontology learning process is to determine which ontologies would best suit a particular problem, hence the selection of an evaluation technique is mandatory. An interesting aspect about evaluation, to an opposed to information retrieval and other areas, is ontologies are not an end product but means to achieving some other tasks. In this sense, an evaluation approach is the fact to also useful to assist users for choose the best ontology that fits their requirements when faced with a multitude of options. In ontology learning process, we can not simply measure how well a system constructs an ontology without raising more questions (Wong et al., 2011). Instead of it begins with some questions about ontology evaluation: is the ontology good enough? If so, with respect to what application? An ontology is made up of different layers, such as terms, concepts, and relations.

3. Pedagogical Domain

Since pedagogical domain is extensive, the aim in this work is just the creation of support tools for the instructors in the classroom. Three topics were researched: learning styles, intelligences types and learning strategies in the class.

3.1 Learning Styles

Learning styles project the way of which a person learn. However, there exist alternatives about how is possible to learn concepts and processing information by humans. Several theories to describe the different types of learning have been proposed in some works. This work adopts as a reference the David Kolb model (Kolb, 1976), where a learning style is determined using the Learning Style Inventory (LSI) scale. The theory proposes a method for describing how students solve problems and apply new knowledge from personal experience within their learning environment. It considers the psychological processes of perception and processing (Olivos et al., 2016). This method comprise four learning styles as follows:

- Active: It includes people who get involved with new experiences, they tend to act first and think the consequences after.
- Reflective: It includes people who are observers and analyze their experiences from different perspectives. They collect and analyze data in detail before take a conclusion.
- Theoretical: People who adapt and integrate their observations into a complex and well logically founded theories. They prioritize logic and rationality before analysis and synthesis.
- Pragmatic: Includes people who test their ideas, theories and new techniques, and try to see if they work in practice. They dislike the long discussions on the same subject. They are practical and attached to reality.

3.2 Intelligences Types

Intelligence is the ability to solve problems, or to create products, that are valued within one or more cultural settings (Gardner, 2001). Humans have a capacities and potentials that can be employed in productive ways (together or separately). This idea originated the multiple intelligences theory. The types of intelligence identified in Gardner (2001) are described below:

- Linguistic intelligence involves sensitivity to spoken and written language, the ability to learn languages, and the capacity to use language to accomplish certain goals.
- Logical-mathematical intelligence consists of the capacity to analyze logically problems, carry out mathematical operations, and investigate issues scientifically.
- Musical intelligence involves skill in the performance, composition, and appreciation of musical patterns. It encompasses the capacity to recognize and compose musical pitches, tones, and rhythms.
- Bodily-kinesthetic intelligence entails the potential of using whole body or parts of the body to solve problems. It is the ability to use mental abilities to coordinate bodily movements.
- Spatial intelligence involves the potential to recognize and use the patterns of wide space and more confined areas.
- Interpersonal intelligence is concerned with the capacity to understand the intentions, motivations and desires of other people. It allows people to work effectively with others.
- Intrapersonal intelligence entails the capacity to understand oneself, to appreciate feelings, fears and motivations.

3.3 Learning Strategies

A learning strategy is a set of procedures that a learner uses consciously, controlled and intentional as flexible tools to learn and solve problems (Barriga & Hernández, 2004). Figure 1 shows the types of strategies published by González and Valle (2011).

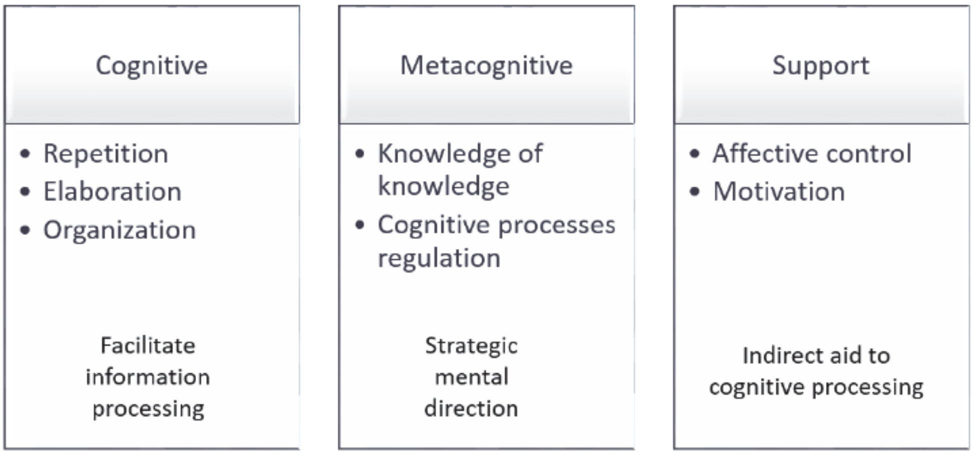

Cognitive
• Repetition
• Elaboration
• Organization
Facilitate information processing
Metacognitive
• Knowledge of knowledge
• Cognitive processes regulation
Strategic mental direction
Support
• Affective control
• Motivation
Indirect aid to cognitive processing

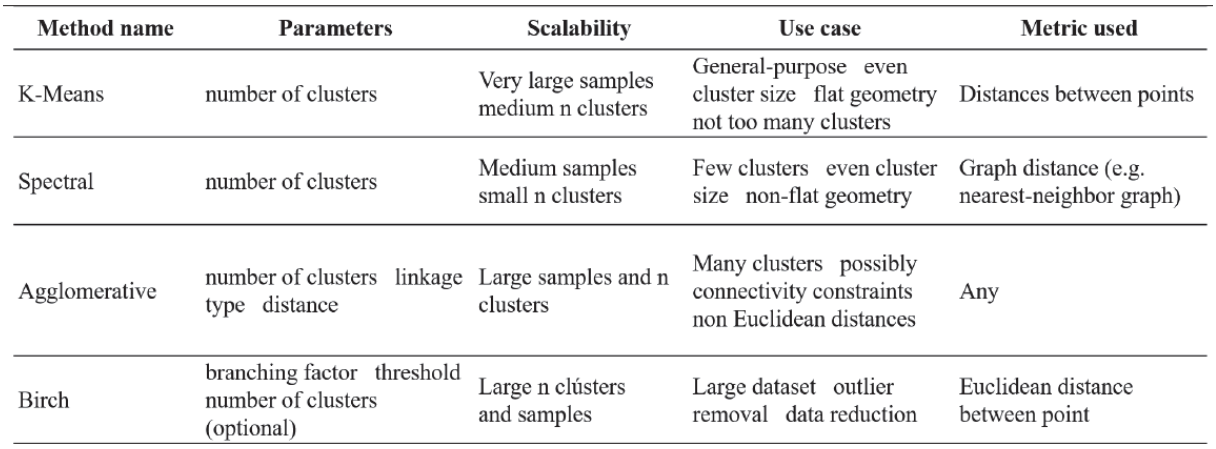
Figure 1. Types of learning strategies

4. Cluster Methods

Clustering is the process of grouping a set of data objects into multiple groups, so that objects within a cluster have high similarity, but are dissimilar to objects in other clusters. Dissimilarities and similarities are assessed based on the attribute values of the objects and distance measures (Han, 2005). Table 1 shows the main characteristics of each cluster algorithm used in this work, and these are explained follows:

- The KMeans algorithm creates subsets of the original input data by trying to separate samples in n groups of equal variance, minimizing a criterion known as the inertia or within sum of squares. This algorithm requires the number of clusters to be specified. It scales well to large number of samples and has been used across a large range of application areas in many different fields (Arthur & Vassilvitskii, 2007).
- The Agglomerative clustering object performs a hierarchical clustering using a bottom up approach: each instance starts in its own cluster, and clusters are successively merged (Pedregosa et al., 2011).
- The Birch method builds a tree called the Characteristic Feature Tree (CFT) for the given data. The data is essentially lossy compressed to a set of Characteristic Feature nodes (CF Nodes). The CF Nodes have a number of subclusters called Characteristic Feature subclusters (CF Subclusters) and these CF Subclusters located in the non-terminal CF Nodes can have CF Nodes as children (Zhang et al., 1996).
- Spectral Clustering defining an affinity matrix between samples, followed by a KMeans in the low dimensional space. Spectral Clustering requires the number of clusters to be specified. It works well for a small number of clusters but is not advised when using many clusters (Luxburg, 2007).

Table 1. Clustering algorithms



Method name	Parameters	Scalability	Use case	Metric used
K-Means	number of clusters	Very large samples medium n clusters	General-purpose even cluster size flat geometry not too many clusters	Distances between points
Spectral	number of clusters	Medium samples small n clusters	Few clusters even cluster size non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Agglomerative	number of clusters linkage type distance	Large samples and n clusters	Many clusters possibly connectivity constraints non Euclidean distances	Any
Birch	branching factor threshold number of clusters (optional)	Large n clústers and samples	Large dataset outlier removal data reduction	Euclidean distance between point

4.1 Evaluation Metrics

Clustering evaluation assesses the feasibility of clustering analysis on a data set and the quality of the results generated by a clustering method. The tasks includes assessing clustering tendency, determining the number of clusters, and measuring clustering quality (Han, 2005). The metrics used are described in the next paragraphs (Pedregosa et al., 2011).

Mutual Information. It is a function that measures the agreement of two assignments, ignoring permutations (labels true and labels predict). The range of results is from 0 to 1, where 1 is the perfect match. The value of the mutual information is not adjusted by chance and will tend to increase as the number of different clusters increases regardless of the actual amount of mutual information between the label assignments. The general formula is defined in 1.

$$MI\,(U,V) = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} P\,(i,j)\, log\left(\frac{P\,(i,j)}{P\,(i)\,P'\,(j)}\right) \tag{1}$$

Where U and V are the labels, actual and predicted by the clustering algorithm, respectively. This concept uses the entropy to perform the calculation of the metric.

Rand Index. It is a function that measures the similarity between the true tags and those obtained by the clustering algorithm; is symmetric, so it does not affect the order in which labels are processed. The range of results goes from -1 to 1, where it negative values are considered bad and 1 is the perfect similarity. If C is a classification task and k is the result of applying a cluster algorithm, *a* and *b* are defined as:

- *a*: The number of pairs of elements that are in the same set in C and in the same set in K.
- *b*: The number of pairs of elements found in different sets in C and in different sets in K.

The Rand index is given by 2.

$$RI = \frac{a+b}{C_2^{n_{samples}}} \tag{2}$$

Where the denominator indicates the total of possible pairs in the data set to be sorted.

Homogeneity. It is the metric in which each group contains only the members of a single class, this concept is complemented by exhaustiveness, in which all members of a given class are assigned to the same group. The values of homogeneity are between 0 and 1 and its mathematical formula is given by 3.

$$h = 1 - \frac{H(C|K)}{H(C)} \tag{3}$$

Where *H* is the conditional entropy of classes.

Fowlkes. This score is defined as the geometric mean of precision and recall 4.

$$FMI = \frac{TP}{\sqrt{(TP+FP)(TP+FN)}} \tag{4}$$

Where *TP*, *FP*, and *FN* are the true positives, false positives and false negatives respectively.

5. Related Work

In the next subsections, some works about pedagogical domain will be presented. Furthermore, some works focused in corpus creation, class extraction and population applied in others domain will be addressed.

5.1 Pedagogical Domain

In Wu (2008) an ontology based on internet education system, which implements the sharing and reusing of learning material in some systems is developed. It is a qualitative research, where an example with a basic computation online course is created describing the system modules: learning, interface and resources.

In Zhu et al. (2008) ENGOnto is presented; the ontology integrates multiple relevant ontologies for personalized agents to deal with dynamic changes of students learning process, interaction between instructor and learning resources in the environment of English language education. The ontology was built manually but authors describe the process of generating knowledge points dependency (class integration). In Zhu and Yao (2009) a learning activity sequencing approach based on ontologies and the learner activity graph in collaborative environments is described. The system is based on ontology, key technologies of the system include: ontology-based knowledge representation and ontology-based knowledge retrieval.

A domain ontology structure who plays an important role for representing higher education concepts and for assist specialized e-learning systems is addressed in Bucos et al. (2010). As the number of classes that are part of the ontology structure and the properties associated to them increases, the ontology was divided into a set of smaller ontologies. Each of these ontologies contains elements that are closely related and they are defined to each other by the properties that bind them together.

An ontology created from CASE diagrams for on-line education is presented in Bagiampou and Kameas (2012); its evaluation is addressed by experts with a manual process. In this work, the focus is on the construction step, where the classes are extracted manually. The ontology creation process from the courses information offered in advanced levels is explained in Ameen et al. (2012), where students can choose courses according with their academic background. Both works present the structure, information, and hierarchy of the classes in a manual way. Other researchers are focused on online education such as Dai and Li (2010), Du et al. (2012) and recently Hssina et al. (2017), where ontologies are manually defined from XML resources available in the Internet, and the evaluation is a manual process too. On the other hand, Hu et al. (2016) proposes an ontology for the internet learning process. In both works is defined an ontology for each entity in the learning process, and the evaluation is conducted with a manual supervised process for domain experts.

There are works such as Uskov et al. (2016) focused on automatic learning; in this paper, an ontology based on the Internet of Things used in a classroom is created, considering the student intelligences. Méndez et al. (2015) proposes to use an ontological modeling for learning personalization that involves students profile according to the multiple intelligence theory by Howard Gardner as well as to use a domain ontology that helps to represent knowledge in virtual learning platforms.

The researches work one or more steps of the ontology learning process using a di_erent pedagogical domain. In Alemán et al. (2017) some researches focused steps of ontology learning process was discussed.

5.2 Corpus Creation

Some researches are focused on the corpus creation, defined for particular domains. Grljevic and Bosnjak (2015) discusses the creation of the relevant linguistic corpus written in Serbian language. The focus is on the sentiment analysis of student generated contents for higher education. In Teixeira et al. (2011) the problem of creating a reference corpus for news classification in fine grained multi-label scenarios was analyzed. The authors propose a semiautomatic approach for creating a reference corpus that uses three auxiliary classification methods: Support Vector Machines, Nearest Neighbor Classifiers and another based on a dictionary.

5.3 Class Extraction

Ontologies for the Use of digital learning Resources and semantic Annotations on Line (OURAL) is presented in Grandbastien et al. (2007), the project includes people from several disciplines (educational science, computer science, and cognitive psychology) building e-learning services. The authors present the extracted class using Natural-language processing (NLP) techniques in unstructured texts about learning situation. Educational domain was also analyzed in Fu et al. (2008), but its application was into Chinese language.

Others works like Ochoa Hernández (2011) present methods for semi-automatic class extraction using a database of Spanish verbs, diathesis alternations and syntactic-semantic schemes (ADESSE tool) (García et al., 2010), where the semantic extracted patterns are the classes. This methodology was applied in educational domain and replicated in financial domain in Ochoa et al. (2011); in both works, the class extraction was completed with the domain expert opinion. A method for class extraction using linguistic patterns and NLP metrics such as morphological labeling is presented in a recent research (Kang et al., 2014).

A data mining approach based on Ontology to classify web documents in order to facilitate applications based on classified text documents like search engines is proposed in Hajiabadi (2014). The ontology is generated by mining Wikipedia. Because of the collaborative efforts of lots of users in adding new articles, Wikipedia expand enormously and consequently it contains almost all of the fields and sub fields. The ontology was named *WikiOnt*, and contains all categories and sub categories exist in Wikipedia.

Theoretical focus that exist inWang (2010) describes two concepts on the SemanticWeb and Ontology and points out the core role of Ontology in Semantic Web. Moreover, a Double-Channel Helix Methodology was described.

6. Methodology

Figure 2 shows the global process for ontology creation (above) and the steps of the proposed methodology in these paper for the class detection process (down). A corpus was built using some academic papers which have two principal characteristics: This papers are focused on social sciences (pedagogy) and written in Spanish language. Besides, papers are related to the principal classes being extracted and joined in an initial corpus. In the last step, an analysis using clustering methods was implemented. Clustering analysis allowed to verify the concepts that the experts considered as the most important for the domain, coincided with the clusters. A Scikit learn tool was used for experiments and its online documentation for results analysis (Pedregosa et al., 2011).

7. Results

The papers from the input test were extracted and preprocessed in plain text. The result of this process is a corpus *A* with 51 instances, where each instance is a paper. This corpus can be described such as A = {*K, T, C*} where:

- K is a paper key and a numeric attribute {1…51}.
- T is the whole paper text, including the title and abstract. In texts, *stopwords* (Note 1), numbers and words with length less to 2 letters were deleted.
- C is the instance class, this is a nominal attribute according to these principal topic in the paper. C =

{*LearningStyle, IntelligenceType, LearningStrategyg*}. Each paper was manually labeled by domain experts according its title and conclusions; the corpus was balanced, thus exists 17 instances for each class.

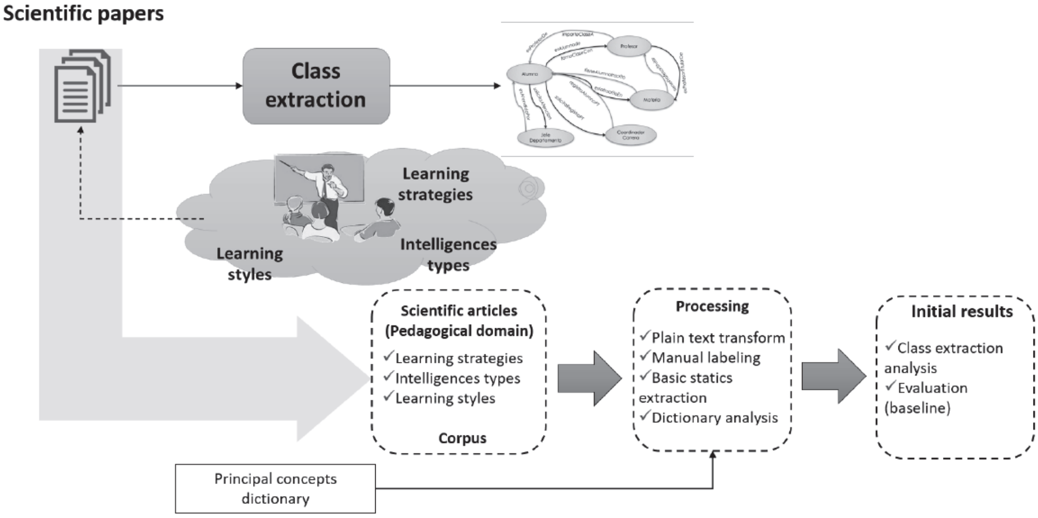
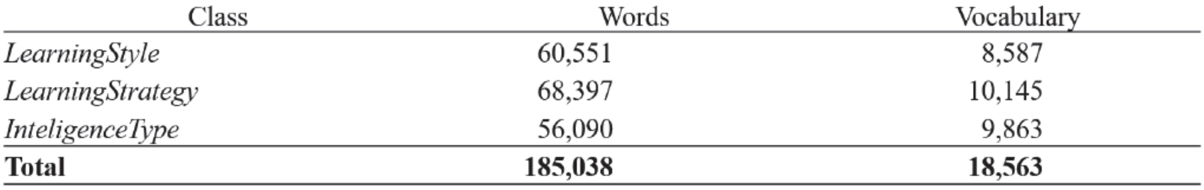


Figure 2. Methodology proposed

Appendix A contains the papers included in the corpus, and Table 2 shows the vocabulary frequency in each class, this analysis was carry out after the initial preprocessing.

Table 2. Corpus Vocabulary



Class	Words	Vocabulary
LearningStyle	60,551	8,587
LearningStrategy	68,397	10,145
InteligenceType	56,090	9,863
Total	**185,038**	**18,563**

LearningStrategy class has more vocabulary, but this class contains two subcategories levels, then, the di_erence between the classes is justified. After the analysis it was concluded that the classes share many words. The final vocabulary corpus contains 18,563 elements.

According to the methodology proposed, a dictionary was built using the principal concepts in each class. First, an initial words list was searched in dictionaries, secondly the concept words and synonyms was added to a dictionary and finally, the *stopwords* were also deleted. The result was a dictionary with 336 words. For example:

- **Initial Word:** Corporal (Type of intelligence)
- **Definition:** la capacidad para utilizar el propio cuerpo para realizar actividades o resolver problemas
- **Synonyms:** morfológico, físico, corpóreo, material
- **Terms Added to Dictionary:** corporal, capacidad, utilizar, propio, cuerpo, realizar, actividades, resolver, problemas, morfológico, físico, corpóreo, material

The cluster methods explained in section 4 were use for the experiments. These cluster were selected because is it possible to determinate the number of clusters as input parameter. In all cases, only this parameter was changed, and the rest of the parameters were established with their default value.

For the experiments, the real corpus classes were represented using numbers 0, 1 and 2 for *LearningStrategy, LearningStyle* and *IntelligenceType* respectively. A word frequency was extracted, as well as, the Term frequency Inverse document frequency (*TF - IDF*) metric (Manning & Schütze, 1999). Besides, two features sets were used: a corpus vocabulary as attributes and the dictionary words as attributes. Table 3 shows the created clusters in each algorithm and the corpus used.

Table 3. Clustering creates in each algorithm and corpus

Corpus	Algorithm	Clusters created
	Real	0 1 0 0 1 0 1 1 2 2 2 2 1 1 0 1 ***0*** 0 2 2 2 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 2 2 2 2 2 2 2 2 2 2
Dictionary	Agglomerative	0 1 0 0 1 0 1 1 1 1 1 1 1 2 0 1 1 0 1 1 1 1 2 2 1 1 2 1 1 1 1 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1
	Birch	0 1 0 0 1 0 1 1 2 2 2 2 1 1 0 1 ***2*** 0 2 2 2 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 2 2 2 2 2 2 2 2 2 2
	K Means	0 1 0 0 2 0 1 1 2 2 2 2 2 1 0 2 2 0 2 2 2 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 2 2 2 2 2 2 2 2 2 2
	Spectral	0 1 1 1 0 1 1 1 1 1 1 1 1 2 1 1 1 0 1 1 1 0 1 1 0 1 2 0 1 1 1 0 1 1 0 0 1 1 1 0 0 0 1 0 1 1 1 1 1 0 1
Vocabulary	Agglomerative	0 0 0 0 0 0 0 0 2 2 1 1 0 0 0 0 0 0 2 1 1 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 1 0 2 1 1 2 0 2 1
	Birch	0 0 0 0 0 0 1 0 2 2 1 1 0 0 0 0 0 0 2 1 1 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 1 0 2 1 1 2 0 2 1
	K Means	0 0 0 0 0 0 0 0 1 1 1 2 0 0 0 0 0 0 1 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 1 1 0 1 0
	Spectral	0 0 0 1 0 0 0 0 0 0 0 0 0 2 0 1 0 0 0 0 0 0 1 0 0 1 2 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 1 0

Greater balance is observed when using the dictionary words as features. Since the initial corpus is balanced in its three classes, one would expect the clustering results to be similar. At this point, a smaller number of attributes generates more balanced groups within the corpus. Table 4 shows application results of the metrics described in subsection 4.1.

Table 4. Clustering metrics

Cluster	Evaluation metric	Dictionary	Vocabulary
Agglomerative	Fowlkes	0.6924	0.5937
	Homogeneity	0.5548	0.4014
	Mutual information	0.5355	0.3762
	Rand Index	0.4861	0.3041
Birch	Fowlkes	**0.9596**	0.5763
	Homogeneity	0.9311	0.3597
	Mutual information	0.9284	0.3329
	Rand Index	0.9406	0.2875
K Means	Fowlkes	0.8445	0.5953
	Homogeneity	0.7774	0.3029
	Mutual information	0.7686	0.2746
	Rand Index	0.7695	0.2661
Spectral	Fowlkes	0.4193	0.4689
	Homogeneity	0.0617	0.0654
	Mutual information	0.0205	0.0231
	Rand Index	0.0066	0.0121

The best results in all metrics were obtained using the dictionary as features and the Birch algorithm, mainly considering the Fowlkes metric (96%). The worst results were obtained using the vocabulary as corpus and the Spectral algorithm, achieving in some metrics less than 5%. Figure 3 presents the best and the worst results obtained.

First, the actual labels (center) are represented with three balanced sets. To the left, the results of the Birch algorithm (dictionary) are shown, which are very similar to the actual tags. For example, in the first group only has one instance less than the actual set, and does not contain any extra element. This group corresponds to the LearningStrategy category. Instance 17 is grouped together with the category of LearningStyle along with all that actually belong to it. This can be seen in Table 3, where the row of real labels and the one corresponding to the birch algorithm with the dictionary are very similar, except for item 17 (highlighted in bold). Finally, the category of IntelligenceType (lower right group in the Venn diagram), is the one that is better defined, since this algorithm does not present overlap with the other two categories.

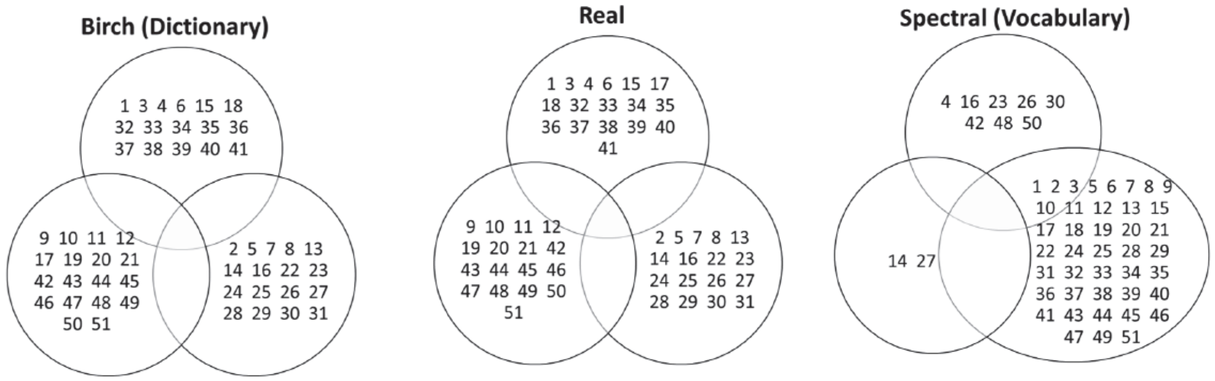


Figure 3. Comparative between real clusters, clusters created by Birch algorithm (dictionary) and clusters created by spectral algorithm (vocabulary). The numbers represents the instances key

Finally, the results obtained with the Spectral algorithm (vocabulary) are shown to the right, where there is an important bias towards a single group, which integrates most of the corpus instances.

With the methodology proposed, a corpus of pedagogical domain labeled manually can be labeled automatically using the Birch clustering algorithm and the dictionary built.

8. Conclusions and Future Work

In this paper, a proposal for a corpus creation for the pedagogical domain was presented, in addition to the use of a main concepts dictionary enriched with synonyms. This corpus was automatically processed and analyzed using clustering techniques, in order to find similarities between the clusters created by established algorithms and the real classes of the corpus.

This proposed is oriented to Spanish language, according to the work state, there are many researches in pedagogy and other domains, but the most of them are in English language and a few in others languages like Chinese. Also, this research involving three principal class and computational techniques for construction and class detection. This processes usually are reported using a manual approach.

As future work, the first tests for the class detection in the corpus will be carried out in order to start the ontological learning process.

Acknowledgments

We would like to thank Vicerrectoria of Research and Graduate Studies, Benemérita Universidad Autónoma de Puebla for supporting this work through the project Model of Teaching-Learning Process applying Ontological Engineering.

References

Alemán, Y., Somodevilla, M., & Vilariño, D. (2017). A proposal for domain ontological learning. *Research in Computing Science, 133*, 63-70.

Ameen, A., Khan, K. U. R., & Rani, B. P. (2012). Creation of ontology in education domain. In 2012 IEEE Fourth International Conference on Technology for Education, (pp. 237-238).

Arthur, D., & Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA ’07, (pp. 1027–1035), Philadelphia, PA, USA. Society for Industrial and Applied Mathematics.

Bagiampou, M., & Kameas, A. (2012). A use case diagrams ontology that can be used as common reference for software engineering education. In 2012 6th IEEE International Conference Intelligent Systems, (pp. 035–040).

Barriga, F., & Hernández, G. (2004). Estrategias docentes para un aprendizaje significativo. Una interpretación constructivista. México: McGraw Hill.

Bucos, M., Dragulescu, B., & Veltan, M. (2010). Designing a semantic web ontology for e-learning in higher education. In 2010 9th international symposium on electronics and telecommunications (ISETC), (pp. 415–

418). IEEE.

Celjuska, D., & Vargas-Vera, D. M. (2004). Semi-automatic population of ontologies from text. In In: Workshop on Data Analysis WDA-2004.

Cimiano, P. (2006). Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Secaucus, NJ, USA: Springer-Verlag New York, Inc. Conference on Multimedia Technology, (pp. 1–4).

Dai, X., & Li, X. (2010). Study of learning source ontology modeling in remote education. In 2010 International

Du, L., Zheng, G., You, B., Bai, L., & Zhang, X. (2012). Research of online education ontology model. In 2012 Fourth International Conference on Computational and Information Sciences, (pp. 780–783).

El-Ansari, A., Beni-Hssane, A., & Saadi, M. (2016). A Multiple Ontologies Based System for Answering Natural Language Questions, (pp. 177–186). Cham: Springer International Publishing.

Faria, C., & Girardi, R. (2014). A domain-independent process for automatic ontology population from text. *Science of Computer Programming, 95*, Part 1, 26-43. Special Issue on Systems Development by Means of Semantic Technologies.

Fu, J., Jia, K., & Xu, J. (2008). Domain ontology learning for question answering system in network education. In 2008 The 9th International Conference for Young Computer Scientists, (pp. 2647-2652).

García, J., Vaamonde, G., & Domínguez, F. G. (2010). Adesse, a database with syntactic and semantic annotation of a corpus of Spanish. In Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., & Tapias, D. (Eds.), Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta. European Language Resources Association (ELRA).

Gardner, H. (2001). Estructuras de la Mente.

González, M., & Valle, A. (2011). Estrategias docentes para un aprendizaje significativo. Una interpretación constructivista. Pamplona: EUNSA.

Grandbastien, M., Azouaou, F., Desmoulins, C., Faerber, R., Leclet, D., & Quenu-Joiron, C. (2007). Sharing an ontology in education: Lessons learnt from the oural project. In Seventh IEEE International Conference on Advanced Learning Technologies (ICALT 2007), (pp. 694–698).

Grljevic, O., & Bosnjak, Z. (2015). Development of Serbian higher education corpus. In 2015 16th IEEE International Symposium on Computational Intelligence and Informatics (CINTI), (pp. 177–181).

Guarino, N., Masolo, C., & Vetere, G. (1999). Ontoseek: content-based access to the web. *IEEE Intelligent Systems and their Applications, 1*4(3), 70–80.

Hajiabadi, H. (2014). Ontology based data mining approach on web documents. *Computer and Information Science, 7*(4), 123.

Han, J. (2005). Data Mining: Concepts and Techniques. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Hssina, B., Bouikhalene, B., & Merbouha, A. (2017). An ontology to assess the performances of learners in an e-learning platform based on semantic web technology: Moodle case study. In Europe and MENA Cooperation Advances in Information and Communication Technologies (pp. 103–112). Springer.

Hu, J., Li, Z., & Xu, B. (2016). An approach of ontology based knowledge base construction for Chinese K12 education. In 2016 First International Conference on Multimedia and Image Processing (ICMIP) (pp. 83–88).

Kang, Y. B., Haghighi, P. D., & Burstein, F. (2014). Cfinder: An intelligent key concept finder from text for ontology development. *Expert Systems with Applications, 41*(9), 4494-4504.

Kolb, D. (1976). Learning style inventory. Boston USA: MA: Hay Group, Hay Resources Direct.

Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing, 17*(4), 395-416.

Manning, C. D. & Schütze, H. (1999). Foundations of statistical natural language processing. Cambridge, MA, USA: MIT Press.

Méndez, N. D. D., Carranza, D. A. O., & Ocampo, M. G. (2015). Representación ontológica de perfiles de estudiantes para la personalización del aprendizaje. *Revista Educación en Ingeniería, 10*(19), 105-115.

Ochoa Hernández, J. L. (2011). Desarrollo de una metodología para la construcción automática de ontologías en español a partir de texto libre. PhD thesis, Departamento de Ingeniería de la información y las comunicaciones. Universidad de Murcia.

Ochoa, J. L., Hernández-Alcaraz, M. L., Almela, A., & Valencia-García, R. (2011). Learning semantic relations from Spanish natural language documents in the financial domain. In Proceedings of the 3rd International Conference on Computer Modeling and Simulation, held at Mumbai, India. Chengdu: Institute of Electrical and Electronics Engineers, Inc, (pp. 104–108).

Olivos, P., Santos, A., Martín, S., Cañas, M., Gómez, E., & Maya, Y. (2016). The relationship between learning styles and motivation to transfer of learning in a vocational training programme. *Suma Psicológica, 23*(1), 25-32.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.

Teixeira, J., Sarmento, L., & Oliveira, E. (2011). Semi-automatic creation of a reference news corpus for fine-grained multi-label scenarios. In 6th Iberian Conference on Information Systems and Technologies (CISTI 2011), (pp. 1–7).

Uskov, V., Pandey, A., Bakken, J. P., & Margapuri, V. S. (2016). Smart engineering education: The ontology of internetof-things applications. In 2016 IEEE Global Engineering Education Conference (EDUCON), (pp. 476–481).

Wang, G. (2010). Methodology research of ontology building in semantic web. *Computer and Information Science, 3*(4), 236.

Wong, W., Liu, W., & Bennamoun, M. (2011). Ontology learning from text: A look back and into the future, *44*, 1-36.

Wu, H. (2008). Research of internet education system based on ontology. In 2008 *Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, 4. (pp. 602-605).

Zhang, T., Ramakrishnan, R., & Livny, M. (1996). Birch: an e_cient data clustering method for very large databases. In *In Proc. of the ACM SIGMOD Intl. Conference on Management of Data SIGMOD,* (pp. 103-114).

Zhu, F., & Yao, N. (2009). Ontology-based learning activity sequencing in personalized education system. In 2009 International Conference on Information Technology and Computer Science, volume 1, (pp. 285–288). IEEE.

Zhu, F., Fok, A. W. P., Ip, H. H. S., & Cao, J. (2008). Engonto: Integrated multiple English learning ontology forpersonalized education. In 2008 International Conference on Computer Science and Software Engineering, volume 5, (pp. 210–213). IEEE.

Appendix A

Papers of initial corpus

1. Macías, María (2002). Las múltiples inteligencias. Psicología desde el Caribe.

2. López, Consuelo (2003). Evaluación de los estilos de aprendizaje en estudiantes de enfermería mediante el cuestionario CHAEA. Enfermería global: Revista electrónica semestral de enfermería.

3. Vargas, Ana (2004). Antes y después de las inteligencias múltiples. Revista electrónica Educare.

4. Guzman, Belkys (2005). Las inteligencias múltiples en el aula de clases. Revista de investigación.

5. García, Hécmy (2007). Variables académicas y estilos de aprendizaje en estudiantes del ciclo de iniciación universitaria. Revista de educación.

6. Enríquez, Álvaro (2007). Estrategias de aprendizaje para la empleabilidad en el mercado del trabajo de profesionales recién egresados. Universitas Psychologica.

7. Santiago, Álvaro (2007). Estrategias y enseñanza-aprendizaje de la lectura. Folios.

8. García, José (2008). Análisis de datos obtenidos a través del cuestionario CHAEA en línea de la página web www.estilosdeaprendizaje.es. Revista de estilos de aprendizaje.

9. Lamas, Héctor (2008). Aprendizaje autorregulado, motivación y rendimiento académico. Liberabit.

10. Paniagua, Liziano (2008). La teoría de las inteligencias múltiples en la práctica docente en educación preescolar. Revista Electrónica Educare.

11. Herrera, Lucía (2009). Estrategias de aprendizaje en estudiantes universitarios. Un aporte a la construcción del Espacio Europeo de Educación Superior. Educación y Educadores.

12. Klimenko, Olena (2009). Aprender cómo aprendo: la enseñanza de estrategias metacognitivas. Educación y Educadores.

13. Esguerra, Gustavo (2010). Estilos de aprendizaje y rendimiento académico en estudiantes de Psicología. Diversitas: Perspectivas en Psicología.

14. Juárez, Jaqueline (2010). Inteligencias Múltiples: Una innovación pedagógica para potenciar el proceso enseñanza aprendizaje. Investigación y Posgrado.

15. Escurra, Luis (2011). Análisis psicométrico del Cuestionario de Honey y Alonso de Estilos de Aprendizaje (Chaea) con los modelos de la Teoría Clásica de los Tests y de Rasch. Persona: Revista de la Facultad de Psicología.

16. Villamizar, Gustavo (2011). Estilos de aprendizaje y rendimiento académico en estudiantes de ingeniería civil. Informes Psicológicos.

17. Juárez, Carlos (2012). El cuestionario de estilos de aprendizaje CHAEA y la escala de estrategias de aprendizaje acra como herramienta potencial para la tutoría académica. Revista de estilos de aprendizaje.

18. Morales, Alejandra (2012). Estilos de aprendizaje en estudiantes universitarios de ingeniería en computación e informática administrativa. Revista de estilos de aprendizaje.

19. Valencia, Nilson (2012). Procesos cognitivos y metacognitivos en la solución de problemas de movimiento de figuras en el plano a través de ambientes computacionales. Tecné, Episteme y Didaxis: TED.

20. Inciarte, Nerylena (2012). Inteligencias múltiples en la formación de investigadores. Multiciencias.

21. Freiberg, Agustín (2013). Cuestionario Honey-Alonso de estilos de aprendizaje: Análisis de sus propiedades Psicométricas en Estudiantes Universitarios. SUMMA psicolÓgica UST.

22. Muñetón, Bahamón (2013). Estilos y estrategias de aprendizaje relacionadas con el logro académico en estudiantes universitarios. Pensamiento Psicológico.

23. Mayora, Isamar (2013). Estrategias Metacognitivas aplicadas en la comprensión de la lectura por estudiantes de Inglés I. Caso Vice. Revista de Investigación.

24. Cala, Ramón (2014). Determinación de los estilos de aprendizaje de estudiantes de 1er curso de ing. industrial y electrónica de la universidad técnica del norte. Ibarra. Ecuador. Revista de estilos de aprendizaje.

25. Juarez, Carlos (2014). Propiedades psicométricas del cuestionario Honey - Alonso de estilos de aprendizaje (CHAEA) en una muestra mexicana. Revista de estilos de aprendizaje.

26. López, Adriana (2014). Estilos de aprendizaje y su transformación a los largo de la trayectoria escolar. Esseñanza e Investigación en Psicología.

27. Salas, Jorge (2014). Estilos de aprendizaje en estudiantes de la Escuela de ciencias del Movimiento Humano y Calidad de Vida, Universidad Nacional. Costa Rica. Revista Electrónica Educare.

28. Sotillo, Juan (2014). El cuestionario CHAEA-junior o cómo diagnosticar el estilo de aprendizaje en alumnos de primaria y secundaria. Revista de estilos de aprendizaje.

29. Campos, Karolina (2014). Actividades de aprendizaje y TIC: Usos entre docentes de la Educación General Básica costarricense. Aproximación diagnóstica. Revista Electrónica Educare.

30. Carrillo, Maria (2014). La teoria de las inteligencias multiples en la enseñanza de las lenguas. Contextos educativos.

31. García, María (2015). Estrategias utilizadas por estudiantes universitarios en el aprendizaje de la lengua extranjera según el género y nivel de competencia. Docencia e Investigación.

32. Sánchez, Iván (2015). Estrategias cognitivas de aprendizaje significativo en estudiantes de tres titulaciones de Ingeniería Civil de la Universidad del Bío-Bío. Paradígma.

33. Vázquez, Ana (2015). La metacognición: Una herramienta para promover un ambiente áulico inclusivo para

estudiantes con discapacidad. Revista Electrónica Educare.

34. Malnieri, Aida (2015). Conocimientos teóricos y estrategias mtodológicas que emplean docentes de primer ciclo en la estimulación de las inteligencias múltiples. Actualidades investigativas en Educación.

35. Nadal, Blanca (2015). Las inteligencias múltiples como una estrategia didáctica para atender a la diversidad y aprovechar el potencial de todos los alumnos. Revista nacional e internacional de educación inclusiva.

36. Ruíz, Daniela (2015). Inteligencias múltiples en alumnos de la Universidad Americana de Asunción. ACADEMO Revista de Investigación en Ciencias Sociales y Humanidades.

37. Sandoval, Aida (2015). Estimación de la inteligencia lingüística-verbal y lógico-matemática según el género y la ubicación geográfica. TELOS. Revista de Estudios Interdisciplinarios en Ciencias Sociales.

38. Campo, Kiara (2016). Metacognición, escritura y rendimiento académico en universitarios de Colombia y Francia. Avances en Psicología Latinoamericana.

39. Hernández, Jaqueline (2016). Metacognición y comprensión oral en L2. Observación de la práctica docente en nivel universitario. Revista electrónica de investigación educativa.

40. Zambrano, Carolina (2016). Autoeficacia, Prácticas de Aprendizaje Autorregulado y Docencia para fomentar el Aprendizaje Autorregulado en un Curso de Ingeniería de Software. Formación universitaria.

41. Álvarez, David (2016). Una mira al futuro ante la relación de las inteligencias múltiples y el rendimiento escolar. Una apuesta hacia nuevas metodologías docentes en la escuela del siglo XXI. Aula de encuentro: Revista de investigación y comunicación de experiencias educativas.

42. Barraza, René (2016). Rendimiento académico y autopercepción de inteligencias múltiples e inteligencia emocional en universitarios de primera generación. Actualidades investigativas en Educación.

43. Cordeiro, Dayane (2016). Múltiples puertas de entrada a la mente de nuestros alumnos: las inteligencias múltiples en el aula de E/LE. Foro de Profesores de E/LE.

44. Garzón, Ana (2016). La i ntegración TIC-Inteligencias múltiples (IM): Una oportunidad de cambio en el proceso educativo. Revista de pedagogía.

45. Perozo, Carmen (2016). Teoría de inteligencias múltiplesuna alternativa en la didáctica de la química. Aula de encuentro: Revista de investigación y comunicación de experiencias educativas.

46. Alducin, JuanManuel (2017). Estilos de aprendizaje, variables sociodemográficas y rendimiento académico en estudiantes de Ingeniería de Edificación. Revista Electrónica Educare.

47. Gómez, Edna (2017). Estilos de aprendizaje en universitarios, modalidad educación a distancia. Revista virtual Universidad Católica de Chile.

48. Díaz, Alejandro (2017). Impacto de un entrenamiento en aprendizaje autorregulado en estudiantes universitarios. Perfiles educativos.

49. Freiberg, Agustín (2017). Estilos, Estrategias y Enfoques de Aprendizaje en Estudiantes Universitarios de Buenos Aires. Psicodebate.

50. Ventura, Ana (2017). Aprendizaje autorregulado en el nivel universitario: Un estudio situado con estudiantes de psicopedagogía de diferentes ciclos académicos. Revista Electrónica Educare.

51. Etchegaray, Maria (2017). Diseño de un recurso multimedia on line basado en Inteligencias Múltiples. Campus virtutales.

Note

Note 1. In computing, stop words are words which are filtered out before or after processing of text. Usually it refers to the most common words in a language.

Copyrights