Automatic Construction of Ontologies for Intelligent E-Learning Systems

Jesús del Peso, Fernando de Arriaga

Abstract—**During the last years a new generation of Intelligent E-Learning Systems (ILS) has emerged with enhanced functionality due, mainly, to influences from Distributed Artificial Intelligence, to the use of cognitive modelling, to the extensive use of the Internet, and to new educational ideas such as the student-centered education, Knowledge Management. The automatic construction of ontologies provide means of automatically updating the knowledge bases of their respective ILS, and of increasing their interoperability and communication among them, sharing the same ontology. The paper presents a new approach, able to produce ontologies from a small number of documents such as those obtained from the Internet, without the assistance of large corpora, by using simple syntactic rules and some semantic information. The method is independent of the natural language used. The use of a multi-agent system increases the flexibility and capability of the method. Although the method can be easily improved, the results so far obtained, are promising.**

***Index Terms*—Ontology, ontology construction, multi-agent system.**

I. Introduction

Ontologies are becoming more important for Knowledge-Based systems. Among them Intelligent E-Learning Systems (ILS) are an extended area, suitable to use ontologies not only for communication purposes among several ILS, sharing different ontologies, but also to enlarge their corresponding Knowledge-Bases.

The new learning revolution, known as E-learning (EL), raised tremendous expectations during the last thirty years. The world of Internet and the web technologies have spread over even in the construction of internal applications running on an intranet. However, once again the actual situation does not match forecasts [1] because in many cases EL has only been a coined name used to describe old applications and ancient optics, with little concern for the learning problem which still lies at the core of the educational process.

To overcome those drawbacks we have followed an approach [2] that aims at the integration of all the elements present in learning, which are: psychological ingredients of learning [3], [4], educational and tutorial aids, and advanced processing techniques to implement all those constituents and manage the whole process. There has been, so far, several attempts and concrete work on this line [5], [6], but there are just a few of them and unfortunately not very well known.

Although some of those systems have already been built by means of agents, a new ILS generation has emerged, characterized by its enhanced functionality and complexity, advanced tutorial functions and cognitive modeling [7], [8], [9]. The introduction of several machine learning techniques, fuzzy logic not only for student's evaluation but also to represent and model the student's and expert's behaviours [10]--[12], and affective computing [13], make possible to design robust and powerful capabilities, and to implement them in a more distributed way, with a better and more efficient integration. The consequences are all related to a closer or deeper insight of the learning process: better and more detailed representation of the domain knowledge, a more realistic design of the student behavior model, personalization of the tutoring aids and advices, and an efficient data processing implementation with learning capabilities. The evaluations so far carried out [14], [15], proof our statements.

The paper deals with the automatic construction of ontologies, as a feature that could increase in great manner the capabilities of ILS. The fact that most of today's ILS rely on multi-agent systems, emphasizes the importance of the automatic or semi-automatic construction of such ontologies, task that can be undertaken by the multi-agent system.

II. Automatic Construction of Ontologies

Several steps can be followed in the automatic construction of ontologies: a) identification and retrieval of concepts (terminology); b) identification and obtainment of taxonomic relationships (relationships of hyponyms [16]; c) identification and extraction of non-taxonomic relationships as, specific relations of interest or relationships of meronyms [17].

Among the main techniques used for extracting information from textual sources we can quote: statistical procedures for natural language processing, by frequently using a *corpus* of documents that includes at least a representative set of environment related documents, and another set of general documents, not limited to the environment of the ontology we want to build; techniques for natural language analysis including morphologic, syntactic and lexical-semantic "shallow" analysis; and finally data mining techniques, even with schemes for automatic learning.

J. del Peso, a Ph.D. candidate, is with the Department of Applied Mathematics, Universidad Politécnica de Madrid, 28040, SPAIN (e-mail: jdelpeso@mat.upm.es).

F. de Arriaga is with the Department of Applied Mathematics, Universidad Politécnica de Madrid, 28040 SPAIN (phone: 34-91-336-7286; fax: 34-91-336-7289; e-mail: farriaga@mat.upm.es).

Statistical techniques are usually followed to identify the ontology terms [18]--[21]. First of all the frequencies of appearance of the concepts within the environment related documents are obtained. Then the same frequencies of appearance are obtained, but using now the set of general documents, not related to the domain. The selection of concepts is done as a function of all these frequencies. Standard measures, such as TF (*term frequency*) or TFIDF (*term frequency-inverted document frequency*) can be used [22] or even more elaborate measures [23].

For the construction of the conceptual hierarchy lexical-syntactic patterns could be used. The are based on heuristic methods using regular expressions for information extraction. The purpose is to define regular expressions to capture recurrent expressions, allowing so to map them into semantic structures [16]. Hierarchical clustering, other statistical method, is also used [24].

For identification of non-taxonomic relationships several methods are currently used: a) lexical-syntactic patterns, usually dependant on the domain [16]; b) semantic patterns consisting on verbal patterns [18] also dependant on the domain; c) data mining techniques for the obtainment of general rules of association [25], [26]. The procedures are based on a statistical analysis that does not allow to know the nature of those relationships; d) other statistical techniques, to obtain pairs of concepts and possible relationships between them, according to the domain based documents [19], [21].

So far, the methods most currently used for the automatic construction of ontologies are statistical, based on large *corpora*, or domain dependant.

Instead, our approach deals with the automatic construction of ontologies, based only on the knowledge of natural language. This knowledge will be fundamentally syntactic. Our purpose is to avoid statistical methods and the construction of large corpora. Besides, we attempt to obtain a general method for extracting ontologic information, independent of the discourse domain. The use of multi-agent systems allows the method to be distributed and scalable, and can be integrated into an information retrieval system.

III. Alternative Approach

The starting point is the collection of a set of documents that will be analyzed by several agents to construct the ontology corresponding to the domain of those documents. These collections could be different for each one of the agents but all of them integrated by documents that are candidates to belong to the chosen domain. But it is also possible that not all of the documents pertain to the same domain or that the documents content could not be homogeneous.

The method is based on the concept of *ontologic network*, that will be defined later on. An ontologic network is, actually, a formal representation of an ontology, able to be constructed and processed in a distributed way by the agents. This representation allows several operations, as aggregation of ontologic networks to obtain a more general network. It allows also the evaluation of the relevance of the components of an ontologic network as well as the application of different functions, as *pruning,* to select elements with real interest from the point of view of their information content.

A. Principles of the system

The basic principles of the system are the following:

a) several agents are dedicated to the construction of the ontology which corresponds to a certain domain;

b) each agent will carry out the analysis of a set of documents, candidates to belong to the domain;

c) each agent will combine the different ontologic networks produced for each of the documents in order to obtain an enhanced network with information relative to the collection of documents analyzed by the agent. This enhanced network will be called *final A_n-agent network*;

d) agents finally interchange their own final networks to elaborate a *global aggregated network*, that includes by aggregation, the ontologic information obtained by all the agents;

e) finally, the global aggregated network will be evaluated and pruned to select the elements with relevance for the searched ontology.

B. Agents processing and coordination

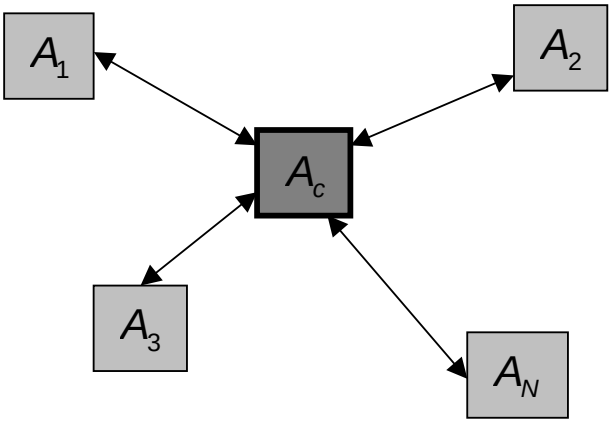

Figure 1: centralized agents control

Agents coordination can be realized in a centralized or decentralized way. The method can be used in combination with any kind of agents coordination, but here, due to space limitation we will restrict ourselves to the centralized way.

In this case there is a *control agent* A_c in charge of: a) controlling the activities of the remainder agents; b) assigning candidate documents to be processed by each agent; c) construct the global final network, as a result of the aggregation of the final networks elaborated by each agent.

That way it will be necessary a first phase for agents initialization. In that phase the control agent assigns domain candidate documents to each agent. Once they have finished up their work the control agent gets the final networks for each agent, and proceeds to construct the global final network. Finally, it selects the relevant elements of the global network to obtain the definite ontologic network.

IV. Formal Model

The method so far exposed is based on a formal approach that lies outside the scope of this paper due to space limitation. But at least we want to outline some definitions and explain the meaning of several operations and functions used by the method.

Definition 1: Simple ontologic network

A *simple ontologic network* on a natural language L, can be defined as an ordered pair

$$O_s = (T_s, R_s)$$

where:

T_s is the set of the network terms, being $T_s \subseteq N_T^L$, N_T^L is the set of terms defined over the natural language L. In general both sets of terms are different; in fact the last one can have an infinite cardinality, however we will only consider finite ontologic networks.

Similarly R_s is the set of relationships among the network terms or equivalently:

$$R_s = \{ r_s = (t_1, t_2, n_{rs}) / \ t_1, t_2 \in T_s , n_{rs} \in N_R^L \}$$

where N_R^L is the set of verbs in the language L that are tagging or giving name to the relations of the network.

Starting out with this definition it is easy to define the concept of extended network.

Definition 2: Extended ontologic network

An *extended ontologic network* on a natural lanuage L can be defined as an ordered triple

$$O = (T, R, M)$$

where:

T is the set of the terms of the network, being

$$T = \{(t_s, k_t) / \ t_s \in T_s, \ , k_t \in N \}$$

with $T_s \subseteq N_T^L$. The integer numbers k_t indicates the number of ontologic networks, precedent of this network, containing the term t_s. R is the set of relations established on the network terms, and M is the number of extended ontologic networks from which O has been obtained. $M = 1$ if O does not proceed from any other, and $M = 0$ for the case of empty ontologic networks.

It is rather obvious to check that every simple ontologic network can be expressed as an extended ontologic network. Similarly, the concepts of sub-networks and equality of extended ontologic networks can be defined.

Now certain operations can be defined on extended ontologic networks, such as reduction, enlargement and aggregation of extended ontologic networks and concepts such as: degree of a term relationship and degree of a generalized relationship. With those elements, the concepts of network connectivity and partition of an ontologic network can be introduced, by using the ideas of ontologic network generated by a term and difference of ontologic networks; it is shown that the partition so defined is unique, therefore a partition function can be obtained.

Several functions have been introduced on the domain of ontologic networks, such as: size of a simple and an extended network, and also the concept of grading the network. For this last purpose several possible grading functions have been defined in order to reach the concept of pruning the network according to a certain grading function.

Finally, several functions for measuring the size of an ontologic network and the similarity of two of them. One of them is described later on, in paragraph VI.

V. Practical Aspects of the Method

A. Construction of ontologic networks from documents

This is the first step of the method. Although it is not necessary, we will assume that for each text an ontologic network will be obtained. Text analysis is mainly syntactic: identification of the most important elements of the ontology. For this purpose the following syntactic elements have been considered: nouns (simple and compound) and noun syntagmas.

In fact only the following principal noun syntagmas have been taken into account: subject and predicate (attribute, direct object, indirect object). It has been assumed that relationships appearing in the network are specified or tagged by the verb they include in the sentence. In consequence this verb will identify the relationship type and its meaning.

As possible interesting relationships, within each phrase, those integrated by terms which are part of the subject, the verb and the predicate terms have been considered. That way each relationship is made of a first term (subject term), a second term (predicate term) and the verb. In case of complex verbs only the principal verb is considered, omitting auxiliary or modal terms. As a first approximation tenses have also been ignored. Ontologies often contain permanent information (non-time-dependent). For this reason it could be necessary to rule out relationships with non-present tenses. In all cases analyzed elements, nouns and verbs, are lemmatized by means of the Porter algorithm [27].

As far as semantic information, the meaning of special verbs, that develop important relationships within ontologies, are considered. So, for example, to be (that gives birth to generalization relationships integrating the ontology taxonomy), and to have (producing relations of meronyms). Verbs denoting spatial relations (and their synonyms) can also be easily introduced.

At this step of the method only terms that take part in relevant or "special" relationships are considered in the analysis. By special relations we mean those relations tagged by a verb whose frequency within the text is greater than a minimum threshold. That way we take into account the effect of real syntactic analyzers and their frequent errors.

B. Construction of the final and global ontologic network

Once all initial ontologic networks nave been obtained, they are combined to obtain the final one according to the following procedure.

First each agent obtains its own final network by using the

operation of aggregation on the set of initial networks it has produced. Then the global network is also obtained by aggregating the final agent networks. Once this global network is produced it is necessary to select the elements of real interest by means of the following procedure: a) grading the global network by means of functions that take into account the terms and relations extensions; b) pruning of the global network after setting up a threshold for the terms grades; c) partition of the pruned network into sub-networks as a result of pruning, but including on each sub-network the elements related with any element of it; d) selection of the most significant sub-network. For this purpose the concept of network size and the functions introduced above for measuring it, are interesting as measure of the information content of a network.

VI. Some Results and Analysis

Several tests have been and are been carrying out. Among them we want to briefly describe some results obtained in the automatic construction of an ontology for "geology".

According to our objectives, a small number of documents was selected (only 70) from the Internet, because most of our Intelligent E-Learning Systems Knowledge Bases are updated with that sort of information. No associated corpus was used to aid this task. The documents were articles related to aspects of the domain of interest. Some other spurious texts were introduced in the set of documents to test the power and capacity of the construction method for selecting the appropriate information. Besides, the domain related documents are not fully homogeneous, in the sense that some of them include out-of-domain elements able to cause errors to ontology construction methods based on pure statistical analysis.

For the syntactic analysis the Stanford Parser [28], provided by the Stanford Natural Language Processing Group, was used. It has to be pointed out that our approach is independent of the natural language used, but it depends on existing parsers if we do not want to get involved in the construction of such a tool.

The rules for terms identification are simple, basically the same introduced in paragraph V *A*, able to identify simple, compound nouns and noun syntagmas acting as subject or object kernels within the analyzed sentences. Anaphoric resolution has not been introduced, although its consideration will improve the efficiency of the described method.

As far as relations identification is concerned, all relations of generalization and of meronyms have been included. For the remainder relations, in order to avoid spurious ones, due to limitations or errors of the parser, a threshold has been used for the appearance frequency of related terms.

As the terms grading function *P(t),* in this example, we have used the following one:

$$P(t) = P^1(t).P^2(t).P^3(t)$$

where:

$$P^1(t) = \alpha.e_T(t)^2 ; \quad P^2(t) = \beta.\sum e_R(t')^2 ; \quad P^3(t) = \gamma.\sum P(t')$$

being $\alpha = \beta = 1$, and $\gamma = 0.1$ This function takes into account for each term not only the square of its own grading value, $e_T(t)^2$, but also the sum of the square of the grading values of all directly related terms, $\sum e_R(t')^2$, $\sum P(t')$. This formula gives more importance to terms close to other terms that have high grades.

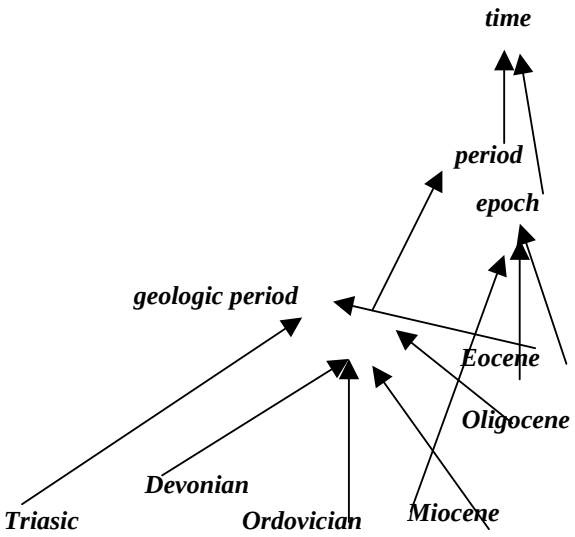

Figure 2. Small part of a taxonomy obtained for "geology"

Many different sets of tests have been carried out looking for distinct objectives. The one we want to, briefly, describe here tries to obtain a taxonomy for *geology*. Several proofs have been carried out with the same set of documents, but changing some parameters. So far the results obtained depend on the threshold value for pruning the aggregate ontologic network. In fact a threshold too low will produce an ontologic network with too many terms and relations, which should not be included.

But in spite of this, and considering the simplicity of the methods here used for information retrieval, the results are promising, and the same conclusion has been obtained in the remainder domains of the tests so far carried out. Choosing an appropriate threshold for pruning, the selection of the terms that integrate the last ontologic network is good. The same thing happens to the set of chosen relations, although here some less relevant relations appear in some more cases. The reason for that could be some deficiencies in the parser or in the grading function that does not consider relations.

Some improvements could easily be introduced in the method; for example, by making more perfect the used parser, or by improving or increasing the semantic information, mostly in relationship to the meaning of verbs conducting the relations identified in the texts. Also the functions and parameters used can be easily refined according to the formal method already developed.

VII. Conclusion

So far, an alternative has been presented to the existing

methods for automatic construction of ontologies. Our approach is based on a formal framework that covers not only the basic definitions of an ontologic network and sub-networks but also the definition of concepts such as

This formal framework constitutes the knowledge that has been introduced into our agents architecture, previously developed, allowing them to carry out in parallel the different tasks for the automatic construction of the ontology.

The method uses syntactic and possibly semantic information that can be extracted from the texts by means of standard tools. But, probably one of its advantages relies on the use of non-statistical methods, allowing in consequence the construction of the ontology based only on a small set of documents, without the assistance of large corpora. Another advantage of the method is to be independent of the natural language used.

The proofs carried out show the practicality and efficiency of the method even with basic syntactic rules, provided the threshold value for pruning elements in the global ontologic network is not very low. Besides, the multi-agent system, working in parallel, can deal simultaneously with a lot of similar problems, increasing that way the system throughput.

The method can be used for many different purposes, such as the automated updating of knowledge bases from the Internet. We are now using this procedure to update knowledge bases of different Neocampus2 spin-off ILS, such as MEDIC2, CentMed2 and Finance.

The method can also be used for automatic information summarization. The elements for the summary could be the homogeneous text sections corresponding to the sub-networks of the partition of the global network obtained after pruning it.

Another possible application of the method could be the categorization of texts or documents. If we have previously obtained a set of different ontologic networks for all possible text categories, it is possible to compare those networks with the one corresponding to the text to categorize. The similarity function developed in our formal frame work would yield the most similar network within the set.

REFERENCES

[1] F. de Arriaga, E-Knowledge Management, E-Learning and E-Commerce: An Evaluation of Their Situation and Tendencies, International Computer Science Institute, Technical Report, Berkeley, 2003, pp. 1-56.

[2] F. de Arriaga, M. El Alami, A. Arriaga, F. Arriaga, J. Arriaga, NEOCAMPUS: Multi-agent Software Environment for On-Line Learning, *Proceedings International Conference on Technology and Education, ICTE'2002*, Badajoz, 2002, pp. 1355-1360.

[3] A.L. Laureano, F. de Arriaga, M. García-Alegre, Cognitive Task Analysis: a Proposal to Model Reactive Behaviors, Journal of Experimental & Theoretical Artificial Intelligence 13 (3), 2001, pp. 227-239.

[4] F. de Arriaga, M. El Alami, A. Ugena, Acceleration of the Transfer of Novices into Experts: The Problem of Decision Making, Proceedings International Conference BITE'2001, Eindhoven, 2001, pp. 157-178.

[5] D. Carmel, S. Markovitch. Learning models of Intelligent Agents. *Proceedings 13th National Conference AAAI'96*, Vol. 2, 1996, pp.137-144.

[6] S. K. Das, J. Fox, D. Elsdon, P. Hammond, Decision making and plan management by autonomous agents. *Proceedings of the International Conference on Autonomous Agents 97*. Marina del Rey, 1997, pp. 276-283.

[7] F. de Arriaga, M. El Alami, A. L. Laureano, "Multi-Agent Simulation for Natural Systems", *Proceedings IASTED International Conference on Modeling and Simulation*, Philadelphia, 1999, pp. 257-268.

[8] A. L. Laureano, F. de Arriaga, "Multi-Agent for Intelligent Tutoring Systems", *Interactive Learning Environments*, Vol. 6, nº 3, 1998, pp. 23-48.

[9] A. Arriaga, M. El Alami, F. de Arriaga, Web-Based Tutors for Collaborative E-Learning, *Advances in Technology-Based Education*, ed: A. Mendez , Badajoz, 2003, pp. 2009-2013.

[10] F. de Arriaga, A. Arriaga, M. El Alami, A.L. Laureano, J. Ramírez, "Fuzzy Logic Applications to Students' Evaluation in Intelligent Learning Systems", *Proceedings II ANIEI International Congress on Informatics and Computing*, Vol. 1, La Paz, 2003, pp. 161-167.

[11] F. de Arriaga, A.L. Laureano, M. El Alami, A. Arriaga, Some Applications of Fuzzy Logic to Intelligent Tutoring Systems, *Proceedings International Conference on Technology and Education, ICTE'2002*, Badajoz, 2002, pp. 1222-1227.

[12] F. de Arriaga, A. Arriaga, M. El Alami, "Fuzzy Intelligent E-Learning Systems: Assessment", *Journal of Advanced Technology on Education,* Vol. 1 (12), 2005, pp. 228-233.

[13] F. de Arriaga, M. El Alami, "Affective Computing and Intelligent E-Learning Systems", *Proceedings IADAT International Conference on Education e-2006,* 2006, pp. 115-120.

[14] F. de Arriaga, M. El Alami, A. Arriaga, "Evaluation of Fuzzy Intelligent Learning Systems", in *Recent Research Developments in Learning Technologies*, Vol. I, ed: A. Méndez, J. Mesa, Formatex, 2005, pp. 109-114.

[15] F. de Arriaga, A. Arriaga, M. El Alami, "Guidelines for the Evaluation of Intelligent E-Learning Systems", in *Technological Advances applied to Theoretical and Practical Teaching*, Iadat, 2005, pp. 142-147.

[16] M. A. Hearst, "Automatic acquisition of hyponyms from large corpora" *Proceedings of he 14th International Conference on Computational Linguistics,* Nantes, 1992, pp. 539-545.

[17] H. Sundblad, Automatic acquisition of hyponyms and meronyms from question corpora, Technical Report SE-58183. Department of Computer and Information Science, Linköping University, Sweden, July 2002.

[18] C. Roux, D. Proux, F. Rechenman, L. Julliard, An Ontology Enrichment method for a pragmatic information extraction system gathering data on genetic interactions, *Proceedings of the ECAI2000 Workshop on Ontology Learning(OL2000)*, Berlin, Germany. August 2000, pp.245-252.

[19] A. Maedche, S. Staab, "Mining ontologies from text". *Proceedings of EKAW-2000,* Lecture Notes in Artificial Intelligence (LNAI-1937) France, Springer, 2000, pp. 59-67.

[20] A. Maedche, S. Staab, "Ontology learning for the semantic web" *IEEE Intelligent Systems,* vol. 16 n. 2, 2001, pp. 72-79.

[21] A. Maedche, R. Volz, "The Ontology extraction and maintenance framework text-to-onto", *Proceedings ICDM Workshop on Integrating Data Mining and Knowledge Management,* San José, 2001, pp. 189-198.

[22] C. D. Manning, H. Schuetze, *Foundations of Statistical Natural Language Processing,* M.I.T. Press, Cambridge, 1999.

[23] R. Navigli, P. Velardi, A. Gangemi, "Ontology learning and its application to automated terminology translation", *IEEE Intelligent Systems,* vol. 18, n. 1, 2003, pp. 22-31.

[24] A. Maedche, V. Pekar, S. Staab, "Ontology learning, Part 1: on discovering taxonomic relations from the web" in *Web Intelligence,* Springer, 2002, Chapter 1.

[25] R. Srikant, R. Agrawal, "Mining generalized association rules" , *Proceedings of VLDB '95,* 1995, pp. 407-419.

[26] A. Maedche, S. Staab, "Discovering conceptual relations from text", *Proceedings of the 14th European Conference on Artificial Intelligence*, Berlin, 2000, pp. 21-25.

[27] M. F. Porter, "An Algorithm for suffix stripping", *Program,* vol. 14, n. 3, 1980, pp. 130-137.

[28] StPar, Java implementation of natural language parsers. Available: http://nlp.stanford.edu/software/lex-parser.shtml