A SEMI-AUTOMATIC BAYESIAN ALGORITHM FOR ONTOLOGY LEARNING

Francesco Colace, Massimo De Santo, Mario Vento
Dipartimento di Ingegneria dell'Informazione ed Ingegneria Elettrica, Università degli Studi di Salerno, Via Ponte Don Melillo 1, 84084, Fisciano (Salerno), Italia

Pasquale Foggia

Dipartimento di Informatica e Sistemistica, Università di Napoli "Federico II", Via Claudio, 21, 80125 Napoli, Italia

Keywords: Bayesian Networks, Ontology, E-Learning

Abstract: The dynamism of the new society forces the professional man to be abreast of technical progress. It is essential to introduce new didactic methodologies based on continuous long-life learning. A good solution can be E-learning. Although distance education environments are able to provide trainees and instructors with cooperative learning atmosphere, where students can share their experiences and teachers guide them in their learning, some problems must be still solved. One of the most important problem to solve is the correct definition of the domain of knowledge (i.e. ontology) related to the various courses. Often teachers are not able to easily formalize in correct way the reference ontology. On the other hand if we want realize some intelligent tutoring system that can help students and teachers during the learning process starting point is the ontology. In addition, the choice of best contents and information for students is closely connect to the ontology. In this paper, we propose a method for learning ontologies used to model a domain in the field of intelligent e-learning systems. This method is based on the use of the formalism of Bayesian networks for representing ontologies, as well as on the use of a learning algorithm that obtains the corresponding probabilistic model starting from the results of the evaluation tests associated with the didactic contents under examination. Finally, we will present an experimental evaluation of the method using data coming from real courses.

1 INTRODUCTION

On-line educational systems represent a rapidly growing research field. Currently, one of the greatest challenges in scientific research is the development of advanced educational systems that are adaptable and intelligent. Methodologies linked to knowledge representation are among the key elements to building intelligent and advanced training systems. In fact, an ensemble of well-structured concepts is able to significantly improve interoperability and information sharing between systems. It can also be efficiently used in intelligent system-supported learning. In literature, such a set of concepts and their relationships describing a knowledge domain is called ontology (Gruber, 1993). It is clear that defining ontologies means formalizing the ways of organizing knowledge, monitoring its transfer procedures and persistence over time, obviously keeping in mind that knowledge mobility will impose continuous changes to the formalized structures. In such a context, ontologies are among the most efficient tools for formalizing knowledge that should then be shared by groups of people (Studer, 1998). Furthermore, it is necessary to identify which items belong to the domain under examination, in order to establish their significance and determine the way in which they relate to the real needs of users. Ontologies have a consolidated reputation as tools capable of satisfying these requests (Swartout, 1997) and, therefore, lend themselves very well for coordinating knowledge organization and distribution in training courses in

Colace F., De Santo M., Vento M. and Foggia P. (2004).
A SEMI-AUTOMATIC BAYESIAN ALGORITHM FOR ONTOLOGY LEARNING.
In *Proceedings of the Sixth International Conference on Enterprise Information Systems*, pages 191-196
DOI: 10.5220/0002625201910196

the field of *e-learning*. Another important and typical aspect of on-line education systems, to which ontologies may surely contribute, is the ability to retrieve the most useful and suitable information to be proposed to students, with the aim of adapting training paths and module sequences to different user needs. The ontology construction process, based on the definition of a graph representing the knowledge domain (the nodes represent the subjects and the arcs represent the pedagogical links), is neither trivial nor easy. Teachers who have to describe the links among the subjects constituting a course often provide a very detailed representation giving birth to ontologies characterized by a large number of states, which could not be easily interpreted and used. A further problem, to which it is difficult to give an unambiguous answer, is related to the evaluation of the links among the different states. As previously said, although direct construction of ontologies is difficult, a source of indirect evidence exists that can be profitably employed for reconstructing "a posteriori" ontology used during a course or a series of lessons: end-of-course evaluation tests. Besides evaluating the students' comprehension of subjects, tests proposed by teachers at the end of a course or a cycle of lessons represent, considering both subject sequencing and propaedeuticity, the ontology really used within the course. The teacher planning the end-of-course evaluation test not only assesses students' level of preparation for the most significant subjects proposed during the lessons, but also tends to describe the ontology outlining the propaedeutic aspects that relate subjects to one another. It may be useful to extract the ontology from these tests, and then evaluate it and refine the propaedeutic relationships among the subjects forming it through the analysis of the answers given by students on such tests. Bayesian networks represent a technique useful for this purpose. Bayesian networks are graph-based probability models where nodes are a set of random variables $\mathbf{X}=\{X_i,\ldots,X_n\}$and arcs represent the causal dependences between variables. In recent years, such networks have been more and more often used for encoding knowledge domains provided by experts with a grade of uncertainty (Heckerman, 2000) and they have proved to be particularly effective for solving data-modelling problems (Conati, 1997). The aim of this paper is to introduce a technique that allow a supervised construction of ontology in order to allow a more easy management of the contents, related to every subject belonging to ontology, by teachers or intelligent tutoring system. In this paper, we firstly define ontologies and the advantages coming from their use in knowledge-based systems. Secondly, we discuss Bayesian networks and how they can easily represent ontology. Finally, we present some results obtained from using Bayesian networks for creating an ontology starting from the answers given by students on tests proposed to them.

2 ONTOLOGIES

Ontologies represent a vast topic that cannot be easily defined, given the disagreements coming from the several methods adopted to build and use them, as well as from the different roles they may play. In 1991, Neches stated that an ontology defines the basic terms and relations comprising the vocabulary of a topic area, as well as the rules for combining terms and relations to define extensions to the vocabulary (Neches, 1991). Later on, Gruber, in the context of knowledge sharing, used the term to refer to an explicit specification of a conceptualization (Gruber,1993). In the field of computer science, ontology represents a tool useful to the learning processes that are typical of artificial intelligence. In fact, the use of ontologies is rapidly growing thanks to the significant functions they are carrying out in information systems, semantic web and knowledge-based systems. The current attention to ontologies paid by the AI community also arises from its recent interest in content theories, an interest that is greater than the one in mechanism theories. In this regard, Chandrasekaran makes a clear distinction between these theories by asserting that, though mechanisms are important since they are proposed as the secret of making intelligent machines, they cannot do much without a good content theory of the domain on which they have to work. Besides, once a good content theory is available, many different mechanisms can be used to implement effective systems, all using essentially the same content. Following this point of view, ontologies are content theories, since their principal contribution consists in identifying specific classes of objects and relations existing in some knowledge domains (Chandrasekaran99). Ontological analysis, therefore, clarifies knowledge structures: given a domain, its ontology represents the heart of any knowledge representation system for that domain. Another reason for creating and developing ontology is the possibility of sharing and reusing knowledge domain among people or *software agents*. It is clear that ontologies are important because they explicate all the possible relations among the concepts belonging to a domain. Once these relations are explained, it will be possible to easily modify them, if our knowledge about that domain changes. These explicit specifications provided by ontologies can also help new users to understand what specific

terms in a domain mean (Uschold, 1992). Furthermore, by offering a unifying structure, ontologies are able to reduce terminological and conceptual ambiguity.

3 BAYESIAN NETWORKS

Bayesian networks have been successfully used to model knowledge under conditions of uncertainty within expert systems, and methods have been developed from data combination and expert system knowledge in order to learn them (Conati, 1997)(Heckerman,1997)(DeCampos,2000). The learning process through Bayesian networks has two important advantages: it is easy to encode knowledge of an expert in a Bayesian network, and such knowledge can be used to improve learning efficiency and accuracy and nodes and arcs of the learnt Bayesian network often correspond to recognizable links and causal relationships. Consequently, it is possible to comprehend and to exploit more easily the knowledge encoded in the representation. A Bayesian network is a graph-based model encoding the joint probability distribution of a set of random variables $\mathbf{X}=\{X_{i,\ldots,}X_n\}$. It consists of a directed acyclic graph S (called *structure*) where each node is associated with one random variable X_i and each arc represents the conditional dependence among the nodes that it joints and a set P of local probability distributions, each of which is associated with a random variable X_i and conditioned by the variables corresponding to the source nodes of the arcs entering the node with which X_i is associated. The lack of an arc between two nodes implies conditional independence. On the contrary, the presence of an arc drawn from the node X_i to the node X_j represents the fact that X_i is considered a direct cause of X_j. Given a structure S and the local probability distributions of each node $p(X_i|Pa_i)$ where Pa_i represents the set of parent nodes of X_i, the joint probability distribution $p(\mathbf{X})$ is obtained from:

$$p(x) = \prod_{i=1}^{n} p(X_i \mid Pa_i)$$

and it is evident that the couple (*S*, P) encodes $p(\mathbf{X})$ unequivocally (Jensen, 1998). In order to construct a Bayesian network for a given set of variables, it is necessary to define some arcs from the causal state to the states that represent their direct effects obtaining a network that accurately describes the conditional independence relations among the variables. Causal semantics of the Bayesian networks has much responsibility for determining their success as knowledge representations in expert systems. Once the network is constructed (through a priori knowledge, or data or a combination of both of them), it is necessary to determine the various probabilities of interest from the model. Such probabilities are not directly stored in the model, it is therefore necessary to calculate them. In general, given a model, the calculation of a probability of interest is known as probabilistic inference (Jensen, 1998).

4 OUR PROPOSAL AND OBTAINED RESULTS

This proposal aims to present a technique able to semi-automatically infer propaedeutic relationships among the different subjects forming a university course. In other words, we intend to define the ontology on which the teacher founds his/her lessons. As previously said, the teacher can have considerable difficulties in delineating relationships among the subjects and their propaedeutic connections. A source of indirect evidence that can be employed for reconstructing a posteriori an ontology used during a course, as well as the propaedeutic connection among the single subjects, is represented by the end-of-course evaluation tests. The teacher planning the end-of-course evaluation tests not only assesses students' level of preparation for the most significant subjects proposed during the lessons, but also tends to describe the ontology on which his/her course was based outlining the propaedeutic aspects that relate subjects to one another. It may be useful to extract the ontology from these tests, and then evaluate it and refine the propaedeutic relationships among the subjects forming it through the analysis of the answers given by students on such tests. In fact, supposing that questions on subjects A and B are posed, and A is considered by the teacher to be a propaedeutic subject to the comprehension of B, it is clear that, in case such a propaedeutic constraint is real, the probability that the student will provide wrong answers to questions relating to B is high if the student gives wrong answers to questions associated with A. On the basis of these considerations, teachers has planned the final test of the first-level course on Computer Science at the Electronical Engineering Faculty of the University of Salerno and the final test of the first-level course on Introduction to Computer Science at the Language Faculty of the University of Salerno. These courses provide first-year students with the foundations of computer science in the first case and introduces to Computer

Science in the second case. At the courses beginning teachers delineated the subjects forming the courses and, then, provide a hypothesis relating to the strength of their relationships. In the case of the second ontology the teacher divided it in two sub-ontology: hardware and software. The result of this process is shown in figure 1 (see appendix). On the basis of the presented ontology, some questionnaires, composed by multiple choice questions, to be filled in by students have been realized. The previously described graph represents the ontologies, but can also be used as a Bayesian network for the inference process. Each node of the networks has two states 'Yes' for complete knowledge of the subject or 'Not' for total ignorance on the subject and represents the probability that a generic learner knows the subject associated with the same node. The student's level of knowledge is evaluated on the basis of the answers given to the questions. The presence of missing values, in other words the state of some variable can not be observable, has not been foreseen. This hypothesis can be obtained imposing that the student must answer to all the questions and thinking wrong a missing answer. Through a Bayesian inference conducted on the previously described networks using a Bayesian inference tool designed and implemented by us, the candidate ontologies networks have learned from data. The inference algorithm used in our tool is the one called "*junction-tree*" introduced by Finn V. Jensen in (Jensen, 1998). For the inferential process we have used data coming from about five hundred questionnaires for the first ontology and three hundred questionnaires for the second and third ones. At this point we have to estimate the strength of propaedeutic relationship between two arguments after the learning of the network. The presence, in fact, of an arc between two nodes in the bayesian nets can be interpreted like a causality relationship between the variables associated to the same nodes so it is important to define a function that is able to evaluate this strength. For the nodes that belong to a bayesian network a good dependence indicator is the cross-entropy function and so defined:

$$C.E.(A,B) = \sum_{a,b} P(a,b) \log \frac{P(a,b)}{P(a)P(b)}$$

where A and B are nodes of the bayesian network and a and b are the states of each node. So according to cross entropy definition we can say that A and B are independent if and only if C.E.(A, B) is equal to 0. However often we have not the real probability distribution of the full network but only an empirical evaluation of it coming from data analysis. So it is incorrect to consider as condition of independence C.E.(A,B) = 0 and we can suppose A independent from B when C.E.(A,B)<e, where e>0 is an arbitrary threshold near to zero. The cross entropy function can also quantify the dependency weight between the nodes. In fact an high value of C.E.(A,B) means a very high preparatory link between the two nodes. In order to suppose that at least the father-child nodes sorting proposed by the teacher is correct we have submitted the data coming from the questionnaires to statistical tests, typical of bayesian network structural learning algorithms, that are able to establish from them the correct father-child nodes arrangement. This tests results have confirmed in the case of the under experimentation ontologies that the arrangement proposed by teachers is correct. So at this point we have set in input to the bayesian network the data coming from the questionnaires in order to obtain the probability values associated to the various states of the nodes. With these values we calculated the cross entropy values among all the single states of the net. Particularly the cross entropy has been calculated both for the arcs proposed by the teacher and for those among brother nodes not signalled. Figure 2, in the appendix, shows the obtained results. On the left side of every figure we can see the cross entropy values for the correct arcs, that represent propaedeutical connection between two topics, while on the rigth side (after the blank column) we can see the cross entropy values for the incorrect arcs. In general we can say the teacher's expected arcs have a greater cross entropy values than other arcs confirming the teacher ontology design. In the case of ontology #1 we have an arc P(8|6) having a cross entropy value in the range of correct arcs. This is not a surprise: in fact in the ontology designing phase teacher had some doubts about this arc. In fact he believed that preparatory links between the nodes 8 and 6 exist but with a low cross entropy value. Instead data show a substantial cross entropy value between these nodes and teacher, according this model, have to refine his ontology proposal.

5 CONCLUSION

In this paper, we have presented a method for learning in a semi-automatical way ontologies representing the didactic contents of an Intelligent Tutoring System and the propaedeutic relationships existing among these contents. In particular, our approach to the problem is based on the use of Bayesian networks. Thanks to their characteristics, these networks can be used to model and evaluate the conditional dependencies

among the nodes of ontology on the basis of the data obtained from student tests. An experimental evaluation of the proposed method has been performed using real student data. The experimentation has demonstrated that the relationships inferred by the system are very similar to the ones that a human expert would have defined, thus confirming the effectiveness of the proposed method. In the future, we aim to integrate the proposed method into a distance learning platform, in order to exploit the inferred ontologies for an adaptive selection of contents. In particular, we intend to use the system to help the teacher in the representation of course reference ontology and in the formulation of tests that provide a better coverage of the course contents, as well as for define per student tests that are automatically adapted to the student training objectives and to his/her level of preparation. This technique can also be applied to the presentation of training contents, thus providing the system with the ability to choose, on the basis of periodical feedback tests, the contents that are most appropriate for each student.

REFERENCES

Gruber T. R., 1993. A translation Approach to Portable Ontology Specifications. *In Knowledge Acquisition, 5(2): 199-220*

Neches R., Fikes R. E., Finin T., Gruber T. R., Senator T., Swartout W. R., 1991. Enabling Technology for Knowledge Sharing. *In AI Magazine, 12(3):36-56*

Chandrasekaran B., Josephson J. R., Benjamins R., 1999. What are ontologies, and why do we need them?. *In IEEE Intelligent Systems, Volume: 14*

Uschold M., Gruninger M., 1992. Ontologies: Principles, Methods and Applications. *In Knowledge Engineering Review, volume 11, number 2*

Heckerman, Geiger, Chickering, 2000. Learning Bayesian Networks: The combination of knowledge and statistical data. *In Machine Learning, vol.4*

Conati, Gertner, VanLehn, Drudzel, 1997. On-Line Student Modeling for Coached Problem Solving Using Bayesian Networks. *In User Modeling: proceedings of the sixth international conference, UM97*

Jensen, F., 1998. An Introduction to Bayesian Networks. *Springer – Verlag, New York.*

Heckerman, 1997. Bayesian Networks for Data Mining. *In Data Mining and Knowledge Discovery 1*

Luis M. de Campos, Juan M. Fernández-Luna, Juan F. Huete, 2000. Building Bayesian Network-Based Information Retrieval Systems. *In proceedings of DEXA Workshop 2000*

Studer R., Benjamins V. R., Fensel D., 1998. Knowledge Engineering: Principles and Methods. In *DKE 25(1-2)*

Swartout B., Patil R., Knight K. and Russ T., 1997. Towards Distributed Use of Large-Scale Ontologies. In *Spring Symposium Series on Ontological Engineering. Stanford University, CA.*

APPENDIX

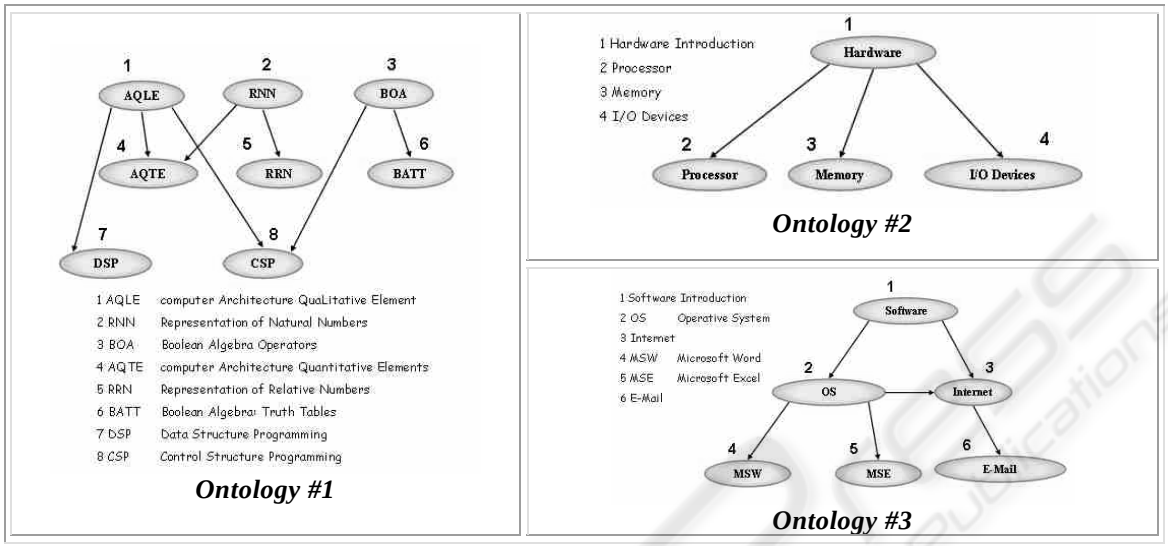

Figure 1: Proposed ontology for the first-level course on Computer Science (Ontology #1) and Introduction to Computer Science (Ontology #2 and Ontology #3)

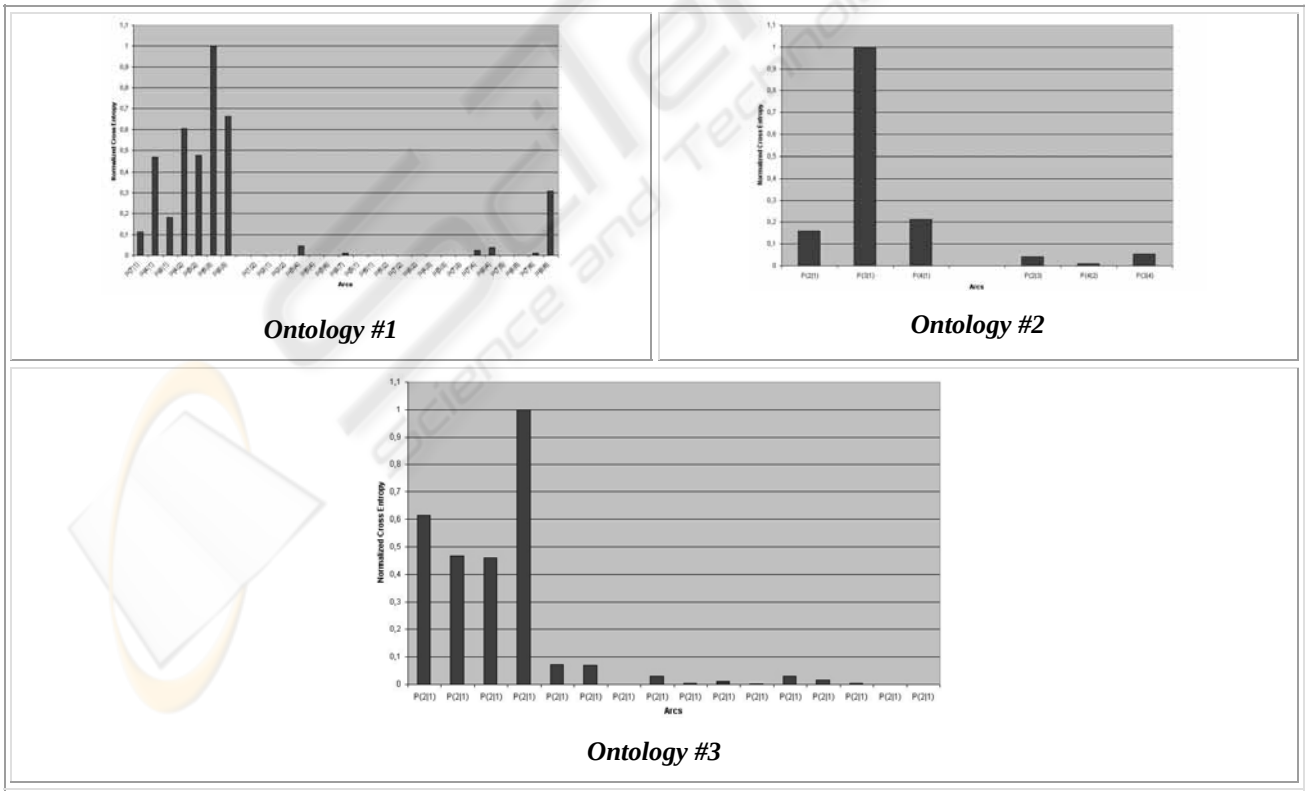

Figure 2: Obtained Results on reference set ontologies