A BAYESIAN APPROACH FOR AUTOMATIC BUILDING LIGHTWEIGHT ONTOLOGIES FOR E-LEARNING ENVIRONMENT

Francesco Colace, Massimo De Santo, Mario Vento
DIIIE, Università degli Studi di Salerno, Via Ponte Don Melillo 1, 84084, Fisciano (Salerno), Italy

Pasquale Foggia
DIS, Università di Napoli "Federico II", Via Claudio, 21, 80125 Napoli, Italy

Keywords: Bayesian Networks, Ontology, MultiExpert System

Abstract: In the last decade the term "Ontology" has become a fashionable word inside the Knowledge Engineering Community. Although there are several methodologies and methods for building ontologies they are not fully mature if we compare them with software and knowledge engineering techniques. In this paper we propose a novel approach for building university curricula ontology through analysis of real data: answers of students to final course tests. In fact teachers design these tests keeping in mind the main topics of course knowledge domain and their semantic relation. The ontology building is accomplished by means of Bayesian Networks.

1 INTRODUCTION

One of the greatest challenges in scientific research is the development of advanced educational systems that are adaptable and intelligent. Methodologies for the knowledge representation are the key elements for building intelligent and advanced training systems. In fact, a set of well-structured concepts can improve interoperability and information sharing between systems. In literature a set of concepts and their relationships is called ontology (Gruber,1993). Ontology is one of the most effective tools for formalizing knowledge shared by groups of people but their building process is neither trivial nor easy but it is very important because it is the starting point of content sequencing both in traditional and on-line courses. Teachers, who have to describe the relationships among the subjects belonging to a course, often provide a very detailed representation creating ontologies with a large number of states that could not be easily interpreted and used. A further problem is related to the evaluation of the links and their semantic values between the different states. In this paper we will propose a method for ontology building that can be applied to knowledge domain related to university curricula. In this case it is more correct to say lightweight ontology because we are finding an advanced taxonomy. In order to solve this problem we have a powerful source of evidence: the end course evaluation tests. Final tests could represent the ontology course because they have been designed by teachers keeping in mind the sequencing and propaedeuticity courses subjects. It may be useful to extract the ontology from answers given by students on such tests. Bayesian networks approach represents an useful technique for this purpose. In recent years, such networks have been more and more often used for encoding knowledge domains provided by experts with a grade of uncertainty and they have proved to be effective for solving data-modelling problems. So the aim of this paper is the introduction of a methodology, based on structural learning Bayesian network algorithms, allowing an unattended lightweight ontology building. So firstly we define ontologies and advantages coming from their use in knowledge-based systems. Secondly, we discuss Bayesian networks and how they can easily map an ontology. In particular we will give some information about structural learning algorithms and their properties. Finally, we will describe the proposed algorithm and we will present some obtained results.

Colace F., De Santo M., Vento M. and Foggia P. (2005).
A BAYESIAN APPROACH FOR AUTOMATIC BUILDING LIGHTWEIGHT ONTOLOGIES FOR E-LEARNING ENVIRONMENT.
In *Proceedings of the Seventh International Conference on Enterprise Information Systems*, pages 386-389
DOI: 10.5220/0002523003860389

2 ONTOLOGIES

The concept of ontology was taken from philosophy where it means a systematic explanation of being. In recent years, however, this concept has been introduced and used in different contexts, thereby playing a predominant role in knowledge engineering and in artificial intelligence. In literature there are many definitions about what an ontology is (Gruber,1993). Ontologies could be represented as a taxonomic trees of conceptualizations: they are general and domain-independent at a superior level, but become more and more specific when one goes down the hierarchy. In other words, when we move from the highest taxonomic levels to the lowest ones, characteristics and aspects typical of the domain under examination are showed. In order to point out this difference in literature we call them heavyweight (deeper ontology) and lightweight (advances taxonomy) ontology respectively. In this paper we will adopt the last one approach keeping in mind this definition of ontology: "An ontology may take a variety of forms, but it will necessarily include a vocabulary of terms and some specification of their meaning. This includes definitions and an indication of how concepts are inter-related which collectively impose a structure on the domain and costrain the possible interpretations of terms"(Uschold,1999). The aim of this paper is to build ontologies, according the previously definition, representing the knowledege domain of university programs.

3 ONTOLOGIES AND BAYESIAN NETWORKS

In this paragraph we will describe bayesian networks and as they can map an ontology. 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. The learning process through Bayesian networks has two important advantages: first of all they easily encode the knowledge of an expert. Secondly nodes and arcs of the learnt Bayesian network represent recognizable links and causal relationships. So user can understand 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 $X = \{X_1, \ldots, X_n\}$. 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 involves conditional independence. On the other hand, the presence of an arc from the node X_i to the node X_j represents 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(**X**) is obtained from: $p(X) = \prod_{i=1}^{n} p(X_i \mid Pa_i)$. In order to construct a Bayesian network for a given set of variables, we need to define some arcs from the causal states to the other ones that represent their direct effects obtaining a network that accurately describes the conditional independence relations among the variables. The aim of this paper is the introduction of an algorithm, based on the formalism of the Bayesian networks, able to infer propedeutical relationships among different subjects (in other terms the ontology) belonging to the knowledge domain of an university curricula. The first step of this algorithm is the introduction of a mapping between Ontology and Bayesian Network. In our ontology model nodes represent the subjects belonging to the course knowledge domain and the arcs mean a propaedeutical relationship among the nodes. We can map this ontology graph in a bayesian network in the following way: the bayesian networks nodes can model the subjects belonging to the course Knowledge Domain and the knowledge of subject by students while arcs in the same way can mean the propaedeutical relationships among the nodes. Given the previous mapping strategy our aim is to define the ontology used by teacher in his/her course. Obviously we must define data type and data set for this approach. As previously said the students answers to the end course evaluation tests represent a source of implicit evidence. In fact, teachers through the end-of-course evaluation tests not only assess students knowledge for every subjects, but describe the course ontology and outline the propaedeutic aspects that relate subjects each other. On the basis of these considerations, teachers have designed 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 Languages Faculty of the University of Salerno. In order to design the reference ontologies teachers used the approach introduced in (Colace, 2004). We must outline that this process was very long and hard for teachers. The result of this process is shown in figure 1. Each node of the networks has two states and shows the probability that a generic learner knows the subject associated with the same node. We have supposed that each node can assume only the following two states (random Bernoullian variable): state 'Yes': complete knowledge of the subject and state 'Not': total ignorance on the subject. The student level of knowledge could be evaluated on the basis of the answers given to the questions (a set of questions is proposed for each subject).

4 AN AUTOMATIC ALGORITHM FOR BUILDING ONTOLOGIES FROM DATA

As previously said our aim is the introduction of an algorithm able to infer automatically propaedeutical relationships between the different subjects forming an university program. In the previous section we defined the general structure of our ontologies and the way to map them in bayesian networks. In this section we will describe our automatic algorithm for building ontologies. The description of the desired automatic algorithm, able to build an ontology from data analysis, could be described in the following steps: to collect data, to collect the nodes of bayesian networks (also ontology nodes) and to learn the structure of ontology (relationships and their strength) through a bayesian statistical inference. In our scenario an effective approach could be the use of structural learning algorithms that can build Bayesian networks (and in our scenario ontologies) using only data. The main aim of structural learning algorithms is to point out the relationships between the entities of a knowledge domain and to specify the causality relationships starting from the observation of domain variables values. More details on structural learning algorithms are in (Neches, 1991). In literature there are many structural learning algorithms but they are not able to achieve good results for every data set and structure. In order to maximize the correct building probability we use a multiexpert approach (Kittler,1998). We selected five structural learning algorithms in order to use them according a majority vote multiexpert approach. The algorithms are: the Bayesian algorithm, K2 algorithm, K3 algorithm, PC algorithm and TPDA algorithm. The main steps of our algorithm are:

- Insert as inputs of every structural learning algorithms bayesian networks nodes and data
- Collect the results (bayesian networks) of every structural learning algorithms and arrange them in a single networks according to a majority vote multiexpert approach. In particular we have an arc between two nodes if and only if three experts say that. The arc sense of direction is obtained in the same way (obviously considering only the experts that point out the arc presence).

We have selected seven networks in order to test the algorithm effectiveness in the building process. In table 1 there is a briefly description of all selected networks and of their related dataset.

Table 1: Analysed Networks.

Network Name	Nodes Number	Arcs Number	Data Set Samples
Alarm	37	46	10.000
Angina	5	5	10.000
Asia	8	8	5.000
College	5	6	10.000
Led	8	8	5.000
Pregnancy	4	3	10.000
Sprinkler	5	5	400

In order to evaluate the performances of algorithm we used this index(Colace, 2004):

$$\text{Global Learning} = \frac{\Sigma \text{ Correctly Oriented Arcs}}{\Sigma \text{ Correctly Oriented Arcs} + \Sigma \text{ Wrongly Oriented Arcs} + \Sigma \text{ Added Arcs} + \Sigma \text{ Missing Arcs}}$$

This index measures the algorithm performance in the learning correct network topology and correct arcs orientation. In Table 2 there are the obtained results of our algorithm compared with the results obtained by best single expert.

Table 2: Obtained results of multiexpert approach versus the results of best expert .

Network	Global Learning Multi Expert	Global Learning Best Single Expert
Asia	1	1
Sprinkler	1	0.83
Alarm	1	0.96
Angina	1	1
Led	0.75	0.55
Pregnancy	1	1
College	0.86	0.67

After this first phase we have used our algorithm on the knowledge domain provideed by teachers as previously described. For the experimentation we have used data coming from about nine hundred questionnaires for the first ontology and seven hundred questionnaires for the other ones.

Table 3: Obtained results of multiexpert approach versus the results of best expert in the real cases.

Network	Global Learning Multi Expert	Global Learning Best Single Expert
Ontology#1	0,50	0,18
Ontology#2	0,80	0,43
Ontology#3	0,57	0,29
Ontology#4	1,00	1,00

Analysing the obtained results (table 3) we can observe as the algorithm offers good results although we have not furnished any type of “a priori” knowledge to the system and a low number of samples that makes worse the performances of structural learning algorithms. In the case of first-level course on Computer Science ontology (figure 2 ontology #1) the system is able to recognize all the links between nodes that the teacher defined "strong". The link that is not recognized has, according to the teacher, the lowest value. The web ontology (figure 2 ontology #4) is built correctly since the number of samples is enough to make reliable and strong the process. Also hardware ontology (figure 2 ontology #2) is built correctly except an arc that, according to the teacher, expresses one of the weakest links inside the net. Finally the ontology Software (figure 2 ontology #2) shows a reverse orientation of an arc and adds two new arcs. The reason for these mistakes is the low number of samples. However, the algorithm offers some satisfactory results from the point of view of the determination of the structure of the net reconstructing all the links defined "strong" by teacher.

5 CONCLUSION

In this paper, we have described a method for automatic learning lightweight ontologies that represent subjects (and their relationships) belonging to a course program knowledge domain. Our approach to problem resolution 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 standard datasets and real data. 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 contents selection.

REFERENCES

Colace, F., De Santo, M., Foggia, P., Vento, M., A Semi utomatic Bayesian Algorithm for Ontology Learning, roceedings of ICEIS 04, Porto, 2004

Gruber, T.R, Translation approach to portable ontology specification, Knowledge Acquisition 5, 1993

Kittler J., Hatef D., Matas J., On Combining Classifiers, IEEE Trans. On PAMI, vol. 20 n. 3, 1998

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

Uschold M., R. Jasper, A Framework for Understanding and Classifying Ontology Applications, IJCAI99 Workshop on Ontologies and Problem Solving Methods, Stockholm, 1999.

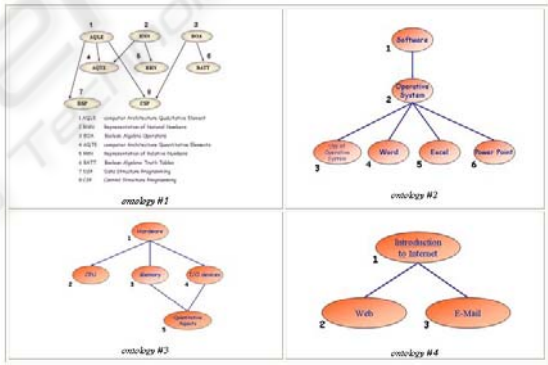

Figure 1: Proposed ontology for the first-level course on Computer Science (Ontology #1) and Introduction to Computer Science (Ontology #2, Ontology #3 and Ontology #4).

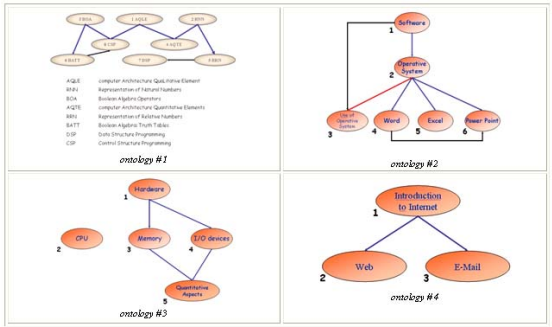

Figure 2: Obtained results. In blue correct arcs, in red wrongly oriented arcs, in black added arcs.