Modelling for Understanding of Scientific Knowledge

G. Albano − G. Gargiulo − S. Salerno

Dipartimento di Ingegneria dell'informazione e Matematica Applicata
Via Ponte Don Melillo− I−84084, Fisciano (Salerno)
{albano,gargiulo}@diima.unisa.it, salerno@unisa.it

1. Introduction

This paper aims at investigating and proposing a new learning method by formalising a representation paradigm for the domain's and learner's knowledge. Indeed, we mean to set forward a representation method of the treated mathematical (and, more generally, formalisable) domain in which additional points are stressed besides the traditional hierarchical, deductive representations. First of all, inductive aspects, considering the problem of finding out which parts of a structure, an example or a problem fit a general framework, therefore going from the particular to the general. This allows, in particular, to thoroughly analyse and overcome misconceptions (often actually built up by direct or socially mediated experience). As a further point we mean to carry out a decomposition into building blocks (of definitions, proofs, exercises etc.) with a point character as close to human experience as possible (e.g., in the sense suggested by modern mereotopology as theory of parts and wholes). Moreover, study of interdisciplinary and infra−disciplinary aspects (links, dependencies, motivations, history etc.) will also be taken into account. Partly as a consequence of this decomposition into blocks, the treated domain is endowed with a richer structure than ordinary textbooks. Moreover, using the graph paradigm, different arcs of several ˆcolours˜ (corresponding to different types of links) may link the same couple of nodes. Another aspect is related to taking into account possible relationships among triples, quadruples etc. of notions and concepts; in fact, some kind of n−ary relationship appears necessary. This will in fact be modelled through the use of supplementary nodes. Our approach will moreover also contribute:

1. to increase the capability and the intelligence of training and learning software environments by building up the framework for a knowledge−level tool able to represent and structure the learner's acquired knowledge. This may be accomplished by using decomposition into informational atomic units and finding the connections among the units themselves: motivational or historical type, difficulty degrees etc. The structure therefore actually consists of both the information units and their links. This, in turn, permits the construction of a tool - based on an abstract environment and an object−oriented approach- which can treat knowledge with mathematical/symbolical instruments, at least in cases (which constitute a significant majority of e.g. University courses) when formalisation and decomposition are possible.
2. to obtain integrated information on learner's preferences, attitudes and flaws, including possible misunderstandings and internal links between different pieces of information, mapping the learner's knowledge and attitudes onto the complete domain knowledge model. This information will thus be used in order to support and improve the learning process. Questions will be chosen in order to obtain maximal information about the learner's knowledge, assuming a suitable closure property of the learner's model. The obtained closed graph model is thus

mapped (imbedded) in the domain's knowledge model and subsequently tested with suitably chosen redundant questions. An optimisation algorithm may then be used to design a path towards the chosen (model) target for learner's knowledge. Effectiveness of an actual process will be tested by actually allowing learners to select targets, e.g. in terms of rating, and testing the results obtained when human experts examine the candidates. Our approach may integrate graph optimisation procedures (question choice, optimal paths) with a collaborative environment, by e.g. automated recording of activity within such a (say, VR−type) environment.

3. To create a knowledge mediation method through a problem−solving approach whose basis relies on the sub−division of notions and exercises into atomic steps. The solving techniques for exercises concerning the domain will be split into atomic steps.
4. To develop tools and methods using a behaviour−based approach in the sense that the model is to be built through the study of reaction to stimuli (input/output: answers to questions and exercises). Once the path to knowledge ripening is started as in 2. above, and one or more informational units have been submitted to the learner, the actual knowledge structure (units/nodes, relations/arcs etc.) is tested and the set of weights used in the optimisation algorithm is possibly dynamically changed according to the actual behaviour of the learner during the learning process. Statistical tests will be used to verify whether the initially chosen set of weights actually corresponds to the average or median case.

2. Basic methodologies and mathematical background

The main methodology that will allow the construction of the domain and learner's models is Graph Theory.

We recall that a *graph* is G defined by an ordered pair (V,E), where V denotes the set of the nodes (or vertices) of G and E denotes the set of the edges connecting pairs of nodes. We define G a *multigraph* if nodes u and v can be connected by parallel edges. A *subgraph* of G is a graph G' whose nodes and edges are all contained in G. A *directed graph* is a graph where an orientation is assigned to the edges. More precisely a directed graph G is defined by the pair (V, A), where V is the set of nodes and A is the set of arcs.

In fact, atomic concepts, notions and techniques - obtained by means of a suitable irreducibility criterion - will be represented as nodes of a directed graph, and links among them are modelled as arcs. Indeed, within the great amount of variations in definitions and names, we use what may be called a *coloured multigraph* where multiple connections between nodes are permitted. Connections may have different ˆcolours˜ and labels according to the degree of difficulty: indeed, we are describing a coloured n−tuple or sequence of ordinary graphs (where at most one line is permitted to join a given couple of nodes). Concerning the construction of the learner's knowledge model, some concepts borrowed partially from geometry and operations research will be used, viz. the concept of ˆextremal˜ points of a graph and of hull (with respect to a certain closure operation). Indeed, in order to minimise the number of test questions, it is assumed that the presence of certain arcs (of e.g. greater degree of difficulty) necessarily entails the existence of certain others. Therefore, a recursive closure operation on graphs may be defined, yielding their hull (with respect to the operation itself). Moreover, a minimal set of ˆextremal˜ nodes and arcs has to be chosen, in such a way to obtain a given target closure. In the geometrical paradigm, this property singles out e.g. vertices of a polygon. Finally, graph−theoretic, operations research algorithms will be

constructed to yield optimal paths from the obtained graph model of the learner's knowledge to the chosen target knowledge.

A common framework for the aforesaid problems is the identification of ˆminimal˜ (according to a given objective function) subgraphs of the learner's model graph for which some special properties hold. It follows that they are computationally ˆdifficult˜ Combinatorial Optimisation problems. For their solution new heuristic algorithms may need to be developed, based on metaheuristic paradigms used in the field of Combinatorial Optimisation, like Greedy Local Search and Tabu Search. Roughly speaking, both methods are based on a ˆcombinatorial˜ extension of the concept of ˆneighbourhood˜ : we start from a seed solution and explore its neighbourhood; if a better solution is found, it replaces the seed solution and the procedure is iterated. They have been successfully applied to many Combinatorial Optimisation problems (Travelling Salesman and Vehicle Routing problems, for instance).

3. Knowledge Domain Model

One of the fundamental steps of the model is the construction of a sufficiently rich structure on the raw set of data, notions and exercises. Indeed, while a traditional textbook's structure is essentially linear, a more ramified type of backbone appears necessary in this context. In fact, the richer the structure, the greater the information that can be recovered from data and feedback - the simplest example of this being a time series as a part of the real line, where both algebraic and order structures play a meaningful role and convey information.

According to this very general idea, we plan to use the graph paradigm. The first step is to choose a suitable level of granularity in splitting the various types of knowledge into atomic parts. An irreducibility criterion appears to be reasonable in the notion joining sense. More precisely, we may define a semigroup structure on notions, where the internal operation is given by joining notions; irreducibility now means that a given notion is a ˆprime˜ i.e. it cannot be expressed (in a non−trivial way) as the product of two or more notions. Of course we may not have a unique factorisation so that, although irreducible elements appear to be well defined, an arbitrary element may be factorised in more than one way into ˆprimes˜ - thus possibly requiring random choices during the factorisation process, or additional nodes and links to account for.

Actually, a coloured multigraph seems a suitable extension of the more traditional (black−and−white and simple) definition of graph. More precisely, different ˆcolours˜ for arcs may account for different types of links between concepts and notions (technical, historical, motivational etc.) so that more than one arc may connect the same couple of nodes; at the same time, each link will be labelled with a degree of difficulty. We remark that this approach is focused on bi−directional links rather than on hierarchical (linear orderings and linear graphs) connections.Since the definition of a graph is by its own nature binary, some extensions of the standard basic ideas in graph theory seem necessary to take into account more general patterns. In fact, in many cases notions and ideas are linked not only in pairs but also in sets of three or more. Therefore, rather than considering an n−term extension of graph theory (hypergraphs), a reduction argument may well fit as follows: a three−term relation is regarded as a binary relation between a single node and a couple of nodes (hence, an arc). This means that supplementary nodes must be added, so as to give the graph the necessary additional dimensions needed to turn a multi−term relation into a two−term one. The main reason for this is that we can exploit the large availability of methods and techniques designed for binary graphs, at the cost of a slight increase in the complexity

in the set of nodes (real and fictitious), and, consequently in a larger amount of memory needed. This seems to be suitable also taking account of the relatively small cardinality of non−binary relations.

4. Learner model

The learner model is the source of all types of information about a typical student. The range of functions and capabilities of the student model varies. Usually a student model is required to keep track of the corresponding student's activities and use this information to give a controller guidance and advise at proper times. There are many approaches to implement a student model. The two main approaches are based on Rule−based system and Semantic−net system respectively. In rule−based systems, the student model is represented as a collection of declarative statements and conditional rules. These statements are used to show in which state−of−knowledge the student is, and the rules specify how these statements are related to each other and also how they can be used for future tutoring scheduling and strategies. This model is rather easier to implement, compared to the semantic−net system, but it has two main drawbacks. The first, and the most important one, is that it is very difficult to maintain domain−independence in this model. It is very difficult to write declarative statements such that they are least dependent on the subject−domain and at the same time are rich in content. The second problem is that the number of rules tends to grow very rapidly and thus dramatically affects the system's performance.

The second approach is based on the semantic net model. The skeleton of this model is a general connected non−cyclic directed graph. This graph hierarchically contains all the subject matters, corresponding to a theme, and their subdivisions at its nodes. One of the important features of this semantic net is a partial domain−independence. The graph is organised in a tree−like fashion. At the root of the net is the main domain's name, for instance Discrete Mathematics.

Each node has an arbitrary number of parent nodes and child nodes. The parent−child relation is of dominance and dependency type. This is obviously a uni−directional relation which specifies how different subject are related and depend on each other. Each relation edge between two nodes has multiple fields which determine the specific details of the corresponding relation and its nature.

For example if the relation is of the *general−to−specific* type or of the *semantically dependent* type. The tree−like structure of the graph shows how different subjects are related in a general−to−specific\parent−child relation. For example, a node representing a topic named *graph circuit* is the parent of the node representing *Hamiltonian circuit* and also of the node representing *Euler circuit* . These topics would make the node *graph circuit* one of their parents. Each node can have an arbitrary number of children and parent node. Each of these parent−child edge carries a *relevance factor* which shows the strength of the corresponding parent−child relation. Of course a strong dependence is represented by a relevance factor equal to one and on the other hand when the relation is very weak, the corresponding factor is much closer to zero than to one.

A fundamental issue concerning the construction of the learner's model is a suitable and possibly optimised choice of questions (including exercises). A possible model that tries to mimic an expert teacher's strategy may be devised as follows. First, an order−based relation is defined so that, once a correct answer establishes a link between concepts, some others (somewhat subordinate to the previous one) are assumed to be present. This recursively yields a hull operator: some nodes and links span a certain set. The choice of test question can now be made in such a way as to obtain a minimal set of nodes and arcs spanning a given target set - just like vertices optimally span a

polyhedron and a base optimally spans a linear space.

A particularly relevant point in the construction of an accurate model of a learner's model, both in a traditional and in non−traditional context, is answer evaluation. A first, widely used, method imposes restrictions to possible answers, such as length restrictions (to reduce ambiguity and decrease answers processing time) or even pre−defined multiple−choice questions. A possible drawback is a corresponding restriction on the type and quantity of information that can be recovered by the analysis of the answers - because of greater significance of guesses and smaller scope of answers. On the other hand, even greater difficulties arise when dealing with automatic interpretation and analysis of natural language; thus, for example, even if a question has a supposedly a unique answer, the lack of a normal form (a standard simplification algorithm that determines whether two elements of a given domain are equal, such as for the word problem for groups or semigroups) may even cause ambiguity in the identification of the answer as correct or incorrect. A possible alternative is the creation of a suitable communication protocol and tools between human expert evaluators and learners. We emphasise that, while this choice is far from making the whole approach trivial, it effectively focuses attention on the learner's knowledge modelling aspects.

Once the global knowledge graph is constructed, the naturally following step relates to its local version, i.e. the learner's knowledge model. The latter is conceived as a subgraph of the former, since only a (generally speaking, proper) part of the arcs and nodes may actually be present. Since a complete direct reconstruction, through questions and exercises, of the learner's subgraph is obviously impracticable, a more subtle strategy is necessary. Indeed, the difficulty level degrees will be used as a linear ordering criterion. More precisely, once a correct answer establishes a link between concepts, all links (within a certain subject) whose degree is not greater than the given link are assumed to be present. The process may now be iterated until no more changes occur (i.e. a ˆfixed point˜ is reached). Incidentally, the previously sketched approach naturally fits the use of such Computer Algebra Systems as Mathematica , where a rule−based evaluation procedure is recursively applied to abstract symbolic expressions until a fixed point is reached.

Besides the more hierarchical matters, it seems necessary to account for logical dependence of concepts and notions belonging to different subjects. Therefore, a supplementary set of arcs − with weights accounting for e.g. strength of logical correlation − will have to be used in order to serve as a model of more general ideas than logical subordination of a concept to another.

While a tree−pruning procedure appears essential in order to construct the learner's model using a reasonable range of questions, some kind of hypothesis testing seems necessary to prevent over− and underestimation of errors. Therefore, some redundancy will be introduced in the extremal set of test questions - whose span optimally covers the knowledge domain. Thus, a random number of randomly chosen test nodes will be chosen at a (possibly randomly chosen) recursion level of the closure procedure (whose final result actually covers the knowledge domain). In case the test nodes (corresponding to correct answers) actually span a subgraph which greatly differs from the parent nodes (belonging to the minimal set), the learner's model construction procedure will restart using fewer recursion levels (hence becoming more conservative). The whole process (questions and compatibility tests) will be iterated until no more changes in the learner's model occur - with a possible upper bound on the number of questions and a *low reliability* warning in case the process is for some reason stopped.

Evaluation is an immediate consequence of the proposed way of constructing the learner's model, since a snapshot of the learner graph will automatically yield this result: a level k knowledge of subject j . Again, a more refined strategy may yield better results; whence, an evaluation matrix (indeed, a tensor with three indices: subject, types and ˆcolour˜ of links) will be constructed, a

statistical analysis of which will give a more accurate description of the learner's knowledge as well as an additional compatibility test. Thus, a correlation matrix (indeed, a still higher−order tensor) will be computed, whose entries are the (e.g. linear) correlation coefficients of the snapshot matrix The latter are (well−known) numbers accounting for the statistical interrelation between high or low marks in a given subject, type etc. and another. Finally, hypothesis testing techniques will yield, within a given level of uncertainty, an evaluation of the consistency of the overall data.

5. Learning support model

In order to devise a learning support strategy, a fundamental step is the choice of a minimum subgraph to be attained. Such a graph may possibly be different according to e.g. the learner's choices and the learning context. The target subgraph will thus be obtained through a recursive procedure, just like the ones sketched above for hierarchical subgraph generation and logical closure generation, now varying such parameters as the number of iterations, the type of arcs and possibly bounding the degree of difficulty.

Knowledge ripening is the process of adding new notions and connections. In order to exploit graph formalism, a combinatorial algorithm will be developed, based on a generalisation of well−known labelling algorithms for shortest path computation in simple graphs.

Weights are assigned to each node and link of the graph. We point out that weights are different in nature from − although connected with− the difficulty degrees introduced before; actually, we may think of them as the time needed to an average learner in order to acquaint with this notion. A starting set of weights is assigned by experts. Then, based on answer evaluation, a fuzzy algorithm will allow redistributing weights. The learning process is thus modelled in a dynamical system fashion in a suitable space of graphs. Analysis of the corresponding time series will yield both a statistical significance test of the possible discrepancies from the starting set of weights, and - in an aggregate form - a possible compatibility test of the actual average−case character of the weights themselves.

Once the adaptive algorithm has determined the set of weights that best describes the learning process, in regard to a given learner, the final set of weights - compared with the average case weights - will yield a quantitative description of the learner's attitudes and preferences (i.e. historical, technical etc.). Thus, the latter may be represented by the final set of weights, while the former may be modelled through the rate of change of the weights.

6. Conclusions

We introduced a possible line of development for knowledge level tool, particularly suitable to mathematical learning modelling and support. The proposed model is based on well founded mathematical theories such as Graph Theory, Operations Research and Statistics, thus appearing reliable; at the same time, the non−trivial use of the latter theories and their synergic combination seem to make the chosen approach useful and capable of interesting future development. This holds true both from the strict point of view of knowledge modelling and from the vantage point of learning support tools development.

7. References

[1] Turns J., Altman C. J., Adams R. − Concept maps for engineering education: a cognitively motivated to supporting variety assessment functions - IEEE Transactions on Education, Vol. 43, n. 2, (2000), pp. 164 - 173

[2] Wenger E. - Artificial Intelligence and Tutoring systems: computational and cognitive approaches to communication of knowledge - Kauffmann (1987)

[3] D'Amore B. - Elementi di didattica della matematica - Pitagora (1999)

[4] Balacheff N., Sutherland R. - Didactical complexity of learning environments - Journal for computers in mathematics learning, n. 4, (1999), pp. 1 -26