Expert Systems with Applications 39 (2012) 8756–8768

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

User behaviour-driven group formation through case-based reasoning and clustering

Mihaela Cocea [a,b,*], George D. Magoulas [b]

[a] School of Computing, University of Portsmouth, Buckingham Building, Lion Terrace, Portsmouth PO1 3HE, United Kingdom
[b] London Knowledge Lab, Birkbeck College, University of London, Malet Street, London WC1E 7HX, United Kingdom

ARTICLE INFO

Keywords:
Case-based reasoning
Clustering
Collaboration
Group formation
Exploratory learning environments

ABSTRACT

Group formation for collaborative learning activities is a complex and time consuming task. Different criteria have been proposed for grouping learners in computer-based systems, such as performance and social characteristics. User behaviour is, however, rarely considered when groups are formed. This paper proposes an approach based on user behaviour that complements the current research on group formation based on different criteria. For this purpose, we propose a synergetic approach based on case-based reasoning and clustering to form groups on the basis of user behaviour. Case based reasoning is used to model user behaviour, while clustering uses the output of the CBR mechanism as criteria for placing learners in relevant clusters. The proposed approach is illustrated using an exploratory learning environment for mathematical generalisation called *eXpresser*.

1. Introduction

Collaborative activities play an important role in teaching and learning. There is an overwhelming body of research from classroom-based education showing that learning in groups enhances pupils' learning by enabling them to learn from each other, e.g. Cohen (1986), Chi, Bassok, Lewis, Reimann, and Glaser (1989), Mastropieri, Scruggs, and Graetz (2003), Rohrbeck, Ginsburg-Block, Fantuzzo, and Miller (2003), Robinson, Schofield, and Steers-Wentzell1 (2005), and Mcmaster, Fuchs, and Fuchs (2006). Although collaborative learning has been proved successful in classroom situations (Brown & Palincsar, 1989; Slavin, 2003), in computer-supported learning environments it does not seem to lead to the same learning benefits. One contributing factor is the way the collaborative groups are formed, as forming efficient groups is very important to ensure an educational benefit from the group interaction (Daradoumis, Guitert, Giménez, Marquès, & Lloret, 2002).

In classroom situations the grouping is most often done by the teacher based on different criteria related to the goals of the collaborative activities. The teachers' decisions about group formation are based on their knowledge of the students – both from educational and social point of view. In computer-based learning this knowledge is used as well and most approaches to group formation consider learners characteristics related to performance and social features, e.g. Gogoulou, Gouli, Boas, Liakou, and Grigoriadou (2007) and Graf and Bekele (2006). Learning performance, however tends to be an aggregated measure of a learner's ability and does not entirely reflect a teacher's knowledge of a particular student, which is more complex and refined than a single indicator. In the context of computer-based environments, more refined knowledge about the learner could be obtained by monitoring the learners' behaviour. In turn, this knowledge about the learners could constitute a more informative basis for the formation of groups.

In this paper we proposed an approach that used the learner's behaviour as the basis for the formation of groups. To this end, we propose a synergetic approach that combines case-based reasoning and clustering to form groups on the basis of user behaviour. The learners' behaviour is modelled using a case-based reasoning approach which provides the input for the clustering approach. Based on the information about the user behaviour, the learners are placed in relevant clusters. The proposed approach is illustrated in the context of an exploratory learning environment for the domain of mathematical generalisation called *eXpresser*.

The rest of the paper is structured as follows. The next section presents previous works related to user modelling, case-based reasoning and clustering. The following section presents the domain of mathematical generalisation and the exploratory learning environment. Section 4 describes the case-based reasoning approach for learner modelling, while Section 5 presents the clustering approach for user behaviour-driven group formation. Section 5

* Corresponding author at: School of Computing, University of Portsmouth, Buckingham Building, Lion Terrace, Portsmouth PO1 3HE, United Kingdom.
E-mail addresses: mihaela.cocea@port.ac.uk (M. Cocea), gmagoulas@dcs.bbk.ac.uk (G.D. Magoulas).
URLs: http://coceam.myweb.port.ac.uk (M. Cocea), http://www.dcs.bbk.ac.uk/~gmagoulas (G.D. Magoulas).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2012.01.205

presents the integrated approach in the context of a classroom situation, Section 6 provides a discussion of the approach and Section 7 concludes the paper.

2. Related work

This section gives an overview of grouping approaches for collaborative activities, with a focus on criteria for grouping. It also covers previous research in the areas of user modelling, case-based reasoning and clustering in the context of exploratory learning environments, and outlines the differences between our proposed approach and previous ones.

Collaborative learning has been extensively researched (e.g., Cohen, 1994; Johnson & Johnson, 1994; Slavin, 1995; Vermette, 1998) and one of the contributing factors to successful collaboration has been identified as the formation of groups in a way that each group member will benefit from the collaborative interaction. The criteria for successful grouping, however, are still not well established, not even for classroom collaborative activities. The tendency is to group students based on their achievement (Macintyre & Ireson, 2002) and form heterogeneous or homogeneous groups with the aim to reduce heterogeneity of learning or of social behaviour (Gregory, 1984). It has been shown that low-achieving students learn more in heterogeneous groups than in homogeneous ones, and that high-achieving students benefit equally from heterogeneous and homogeneous groups (Webb, Baxter, & Thompson, 1997). Besides these findings, there is little known about the influence of group formation on the collaborative processes and performance (Webb et al., 1997; Leornard, 2001).

In computer-supported learning, the criteria used for group formation are learners' characteristics related to their performance and social characteristics (e.g. Gogoulou et al., 2007; Graf & Bekele, 2006). A high-level description of the processes involved in group formation was proposed for virtual learning environments where learners do not know each other and several parameters that influence group collaboration were identified: (a) individual and group learning and social goals; (b) relationships among group members; (c) the interaction process; and (d) members' specific characteristics (Daradoumis et al., 2002). To ensure formation of optimal groups, dynamic grouping supported by wireless handhelds has been proposed for classroom use, allowing reconfigurations of groups to find optimal ones (Zurita, Nussbaum, & Salinas, 2005). Consequently, current research indicates that group formation is a complex problem with no straightforward answer.

Our research extends the current research on grouping for collaborative learning activities by taking into consideration the user behaviour in the grouping approach. This is especially important in exploratory learning due to the difference in how the learners interact with exploratory learning systems compared with tutoring systems.

Exploratory learning environments (ELEs), also referred to as open/inquiry/discovery learning environments (e.g. SimQuest (van Joolingen, King, & de Jong, 1997), Adaptive Coach for Exploration (ACE) (Bunt & Conati, 2003), Vectors in Physics and Mathematics (Grigoriadou et al., 1999)) are characterised by freedom, allowing learners to explore the domain rather than guide their learning in a structured manner. These environments are fundamentally different from Intelligent Tutoring Systems (ITSs) (e.g. Algebra Tutor (Koedinger, Anderson, Hadley, & Mark, 1997), Geometry Tutor (Matsuda & VanLehn, 2005), ActiveMath (Melis, Büdenbender, Goguadze, Libbrecht, & Ullrich, 2003)) which are characterised by structured learning.

The differences between ITSs and ELEs in how learners interact with the system leads to differences between requirements for user modelling for the two types of systems. A typical ITS has a domain model and a student model, where the student model is a "copy" of the domain model with attached information about how much the student has learned/covered from that domain model and to what degree; typically, the domain consists of concepts and/or common misunderstandings (or bugs). Such an approach is not entirely suitable for ELEs because ELEs are typically built for exploring the so-called ill-defined domains (Lynch, Ashley, Aleven, & Pinkwart, 2006), which are characterised by more complex problems that often have multiple solutions, for which a complete domain model is very difficult to develop.

Previous attempts at learner modelling for ELEs include: (a) the use of heuristics to guide the learning process in a physics domain (Veermans, 2003); (b) Bayesian networks in a mathematical functions domain (Bunt & Conati, 2003); (c) neuro-fuzzy systems for student diagnosis in a physics domain (Stathacopoulou, Magoulas, Grigoriadou, & Samarakou, 2005); (d) Fuzzy sets for modelling cognitive states in a computer-based learning environment for Newtonian dynamics (Andaloro & Bellomonte, 1998); (e) eye-tracking for modelling meta-cognitive characteristics such as self-explanation (Merten & Conati, 2006; Conati & Merten, 2007); (f) a Dynamic Decision Network approach for a dynamic Learner Model allowing reasoning about the learners behaviour and interventions across time (Ting & Phon-Amnuaisuk, 2009).

The systems involved in the learner modelling approaches mentioned above are virtual labs or simulation-based environments in which the students are asked to tune variables until they find the right values. Our system, however, asks learners to construct patterns, and to identify variables and relations, which can be done by following various approaches that we call *strategies* (a detailed description of the system is given in the following section). Thus, learner's actions are more important in our context than concepts. Moreover, our learner modelling component was developed in parallel with the design of the system and of the educational activities, while the Learner Models mentioned previously were developed after the systems were completed and in use. Therefore, these two aspects required a different modelling approach that would address our need to model complex problems with multiple solutions and to provide flexibility and extendability to allow alignment with the development of the system. Consequently, a case-based reasoning approach was used due to its capability to handle both aspects outlined above.

The existing approaches have a number of strengths and limitations that are outlined in the following. Bayesian Networks are an established modelling technique in the context of intelligent tutoring systems; however, it does not fit our purpose as we are not dealing with concepts, but with learner actions. The neuro-fuzzy approach has the advantage of dealing well with uncertainty and of mimicking teachers' reasoning; again, concepts in the form of fuzzy variables are used which do not apply to our situation. Fuzzy sets also deal well with uncertainty, but have a similar drawback – they keep track of the overall knowledge, while we want to keep track of the models for each task in a more detailed manner than a number. The dynamic decision networks have the advantage of dealing with the temporal aspect; however, they are used for modelling skills, while we're interested in modelling knowledge. Moreover, none of these approaches address the need for diagnosing the learner during a task rather than at the end of it, which the CBR-approach can do.

Although CBR has been successfully used in applications for domains like legal reasoning (Aleven, 2003), stock market prediction (Chun & Park, 2005), recommender systems (Kumar, Gopalan, & Sridhar, 2005), and other areas, there is little research on using CBR for e-Learning environments. For example, Han, Lee, and Jo (2005) use CBR in the learner modelling process and call this approach case-based student modelling, while Huang, Huang, and Chen (2007) use CBR and genetic algorithms to construct an

optimal learning path for each learner. CBR is used also in Stottler and Ramachandran (1999) within a case-based instruction scenario rather than a method for learner modelling. We have not found any references in the literature to ELEs that use CBR or CBR combined with other intelligent methods.

The advantage of CBR for learning environments and especially for ELEs is that the system does not rely only on the general knowledge of a domain, but it can also use specific knowledge previously experienced (Han et al., 2005). It also seems promising for improving the effectiveness of complex and unstructured decision making (Huang et al., 2007), especially in combination with soft computing methods.

To model the user behaviour we used a modified version of case-based representation that allowed us to represent strategies (or composite cases) as series of simple cases with certain relations between them. The typical CBR representation involves a case that has two parts: a problem and a solution. When a new problem is encountered, it is matched to the problem part of the cases in the knowledge base. If a good match is found, the solution of the matching case is applied to the new problem; if no good match is found, the solution may be adapted for the new problem and then stored for future use. We, however, are dealing with multiple solutions of the same problem and the aim is to identify which of the solutions is used. Therefore, our composite cases, i.e. strategies, do not have a problem part, as this does not vary. Moreover, as learners often get stuck before finishing their constructions and it is important to identify what the learners are doing *while* they solve the task and not only at the end, the strategies were defined as a series of simple cases rather than a single entity that would not allow inspection of its parts.

In addition, modelling learner's strategies, rather than concepts for example, gives the advantage of having a more holistic view of the learner's perspective of a particular task. In other words, a strategy contains more information than a probability attached to a concept (a common approach for Intelligent Tutoring Systems). Also, this is more appropriate for ill-defined domains (Lynch et al., 2006) for which exploratory learning is more suitable than tutoring, as they are often characterised by complex problems, in which a concept cannot be explored in separation from other ones because the essence lies in the relation between concepts.

The user behaviour modelled by the CBR approach is used in the grouping approach by considering the pedagogical aims of the collaborative learning activities, which in our system (and other ELEs) is to discuss similarities and differences between various approaches in solving the same task. Therefore, two aspects were identified as relevant as criteria for grouping: the strategy or strategies followed by each learner and the similarity between the different strategies of a task. Consequently, we developed a clustering approach that takes into consideration these criteria.

Clustering methods are used in a variety of domains such as image segmentation (Bong & Rajeswari, 2011), natural language processing (Ushioda & Kawasaki, 1996), wireless sensor networks (Abbasi & Younis, 2007), galaxy formation (Ross, Tojeiro, & Percival, 2011) and gene expression analysis (Dhiraj, Rath, & Pandey, 2009). However, in our searches of the literature we did not find any research in the area of exploratory learning environments that use clustering methods, not any research about grouping for collaborative activities in ELEs.

The next section introduces the domain of mathematical generalisation and *eXpresser*, the exploratory learning environment for this domain. An outline of the difficulties encountered by students in learning mathematical generalisation and how *eXpresser* addresses these issues is presented, and the aims of the collaborative learning activities are also given.

3. Exploratory learning of mathematical generalisation

Mathematical generalisation is at the centre of algebraic expressions, as "algebra is, in one sense, the language of generalisation of quantity. It provides experience of, and a language for, expressing generality, manipulating generality, and reasoning about generality" (Mason, 2002, p. 105). This relation, however, and the idea of recognising and analysing patterns and articulating structure, seems to be elusive to students who fail to understand algebra and its purpose (Geraniou, Mavrikis, Hoyles, & Noss, 2008). Students are unable to express a general pattern or relationship in natural language or in algebraic form (Hoyles & Küchemann, 2002).

Students, however, are able to identify and predict patterns (Mason, 2002) and there are claims that it is not the generalisation problems that are causing difficulties to students, but the way these are presented and the limitations of the teaching approaches used (Moss & Beatty, 2006): "generalising problems are usually presented as numeric or geometric sequences, and typically ask students to predict the number of elements in any position in the sequence and to articulate that as a rule" (Moss & Beatty, 2006, p. 443). A common strategy is "the construction of a table of values from which a closed-form formula is extracted and checked with one or two examples" (Bednarz, Kieran, & Lee, 1991, p. 7), introducing a tendency towards pattern spotting and emphasising its numerical aspect (Noss, Healy, & Hoyles, 1997; Noss & Hoyles, 1996). This approach obscures the variables involved, "which severely limits students ability to conceptualize the functional relationship between variables, explain and justify the rules that they find, and use the rules in a meaningful way for problem solving" (Moss & Beatty, 2006, p. 444).

Another approach that affects students' understanding of generalisation is the focus on mathematical products rather than mathematical processes (Warren & Cooper, 2008; Malara & Navarra, 2003). Malara and Navarra (2003) argue that students should be taught to distance themselves from the result and the operations needed to obtain that result, and to reach a higher level of thinking by focusing on the structure of a problem.

Another issue encountered in teaching mathematical generalisation is the students' difficulty to use letters that stand for the unknown (Küchemann, 1991) and to realise that letters represent values (Duke & Graham, 2007). Secondary school students also tend to lack a mathematical vocabulary for expressing generality (Geraniou et al., 2008) and research reports on students' lack of precision in written responses (Warren & Cooper, 2008).

Taking these aspects into account, a system for teaching mathematical generalisation was developed using an iterative process that involved designing with students and teachers. The main aim was to develop an environment that provides the students with the means for expressing generality rather than considering special cases or spotting pattern (Geraniou, Mavrikis, Hoyles, & Noss, 2009a). Five critical ideas have informed the design of the system: (a) providing a rationale for generality, (b) supporting simultaneously model construction and analysis, (c) scaffolding the route from numbers to variables, (d) working on a specific case 'with an eye' on the general and (e) reflecting on derived expressions (Geraniou, Mavrikis, Kahn, Hoyles, & Noss, 2009b; Noss et al., 2009).

The system, called *eXpresser*, enables construction of patterns, creating dependencies between them, naming properties of patterns and creating algebraic-like rules with either names or numbers. The system is intended for 11–14 year olds and classroom use in UK secondary schools, and it follows the UK curriculum. Individual tasks in *eXpresser* involve building a construction and deriving a rule form it; collaborative tasks involve discussion on similarities and differences between individual constructions and rules.

Fig. 1 illustrates the system, the *properties list* of a pattern (linked to another one) and an example of a rule. The screenshot on the left includes two windows: (a) the students' world, where the students build their constructions and (b) the general world that displays the same construction with a different value for the variable (s) involved in the task (placed in the area 'I need to vary' in both worlds), and where students can check the generality of their construction by animating their patterns (using the Play buttons). The presence of these two spaces allows learners to work on a specific case (in the students' world) 'with an eye' on the general (in the general world).

We illustrate here a task called 'pond tiling' (see Fig. 1) displayed in the students' world with a 4 by 3 blue (darker colour) pond and in the general world with a 9 by 7 pond; the task requires to surround the pond and find a general rule for the number of tiles needed for this purpose. The model for this task can be built in several ways that we call *strategies*. Here we illustrate the so-called 'H' strategy, named after the shape of the construction. The components of this strategy are highlighted in the students' world for ease of visualisation: the 4 by 3 pond, 2 horizontal green (lighter colour) rows of 4 tiles and 2 vertical green bars of 5 tiles.

To provide a rational for generality, the tasks are presented dynamically. For example, the 'pond tiling' task is presented with the image displayed in Fig. 1 in the student's world without any highlighting of structure; this image changes regularly to show different instances of the pattern. This dynamic presentation provides "a rationale for deriving a rule that outputs the number of green tiles for any instance of the pattern, i.e. a 'general' rule giving concrete instantiation to the meaning of 'any' " (Geraniou et al., 2009b, p. 52).

The property list of one of the horizontal bars is displayed in the top right screenshot. The first property (Ⓐ) specifies the number of iterations of the building-block, i.e. the basic unit of a pattern, which is displayed as an icon; the value for this attribute is set to the value of the width of the pond by using a T-box (that includes a name and a value); by using a T-box, the two (or more) properties are made dependent, i.e. when the value in the T-box changes in one property, it also changes in the other one (s). The next properties are *move-right* (Ⓑ), which is set to 1, and *move-down* (Ⓒ), which is set to 0. The last property (Ⓓ) establishes the number for colouring all the tiles in the pattern – for this simple pattern the value is the same as the iterations and is also related to the width of the pond through the use of a T-box. The bottom right screenshot displays a rule for the number of green tiles: $(h + 2) \times 2 + w \times 2$, where h and w stand for the T-boxes in the area 'I need to vary' (the same as the ones in property lists); a T-box can be displayed with name only, value only or both, thus enabling use of multiple representations.

The use of T-boxes helps learners develop multiple representations which is considered beneficial as it leads learners to deeper knowledge acquisition of a domain, that in turn, could lead to knowledge transfer in other learning situations (van der Meij & de Jong, 2006). Also, "having to make the mental transference between representations ... forces reflection beyond the boundaries and details of the first representation and an anticipation of correspondences in the second. The deeper level of cognitive processing can reveal glitches that might otherwise have been missed" (Petre, Blackwell, & Green, 1998, p. 474).

Multiple External Representations (MERs) (as opposed to mental representations) have several functions: to complement, constrain and construct (Ainsworth, 1999). "The first function is to use representations that contain complementary information or support complementary cognitive processes. In the second, one representation is used to constrain possible (mis)interpretations in the use of another. Finally, MERs can be used to encourage learners to construct a deeper understanding of a situation" (Ainsworth, 1999, p. 134).

Through their multiple representation, the T-boxes are scaffolding the route from numbers to variables, emphasising the idea that variables represent values, but those values do not need to be known – hence the enabled display of a T-box with value only, name only or both. Thus the transition from a specific value to a value that also has a name to a name only (i.e. variable) is facilitated: "this stands in contrast to the standard approach in which generalisations are constructed from special cases, and the path to the variable 'n' appears as a separate (often nonnegotiable) cognitive leap" (Geraniou et al., 2009b, p. 54).

To make a construction general, T-boxes are needed to link the different parts of the construction. Without these links, a construction is specific, i.e. it is valid only as a particular instance of the task pattern; a construction can also have some links in place, while others are missing, i.e. the construction is partially general. These concepts of specific, partially general and general are pedagogically important, as they signal the learners' progress in solving the task.

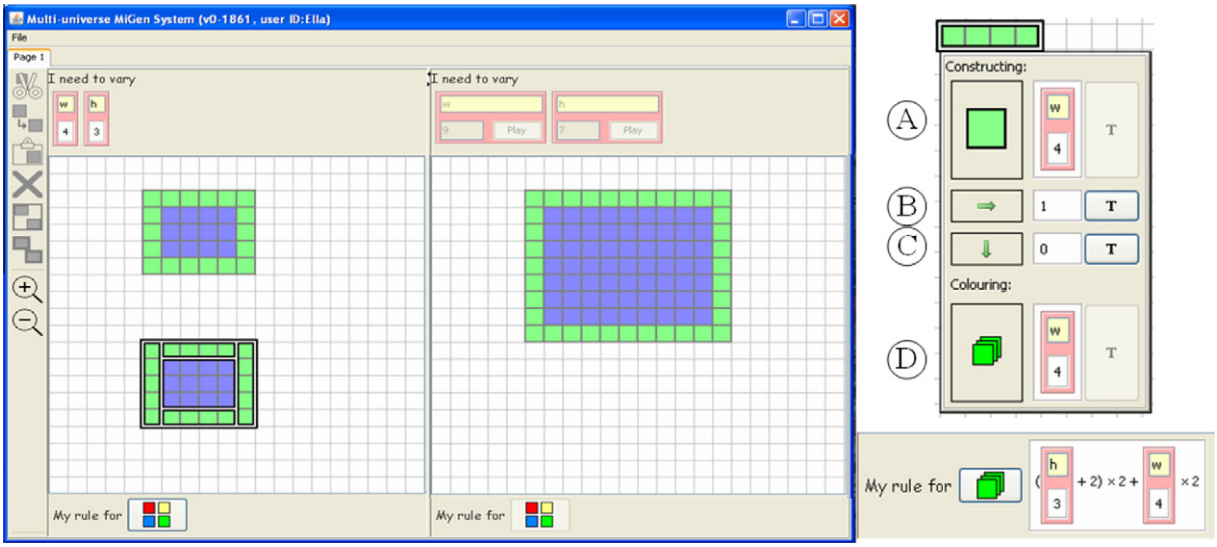

Fig. 1. *eXpresser* screenshots. The screenshot on the left includes a toolbar, the students' world and the general world. The screenshot on the top right shows the property list of a pattern. The bottom right screenshot illustrates a rule.

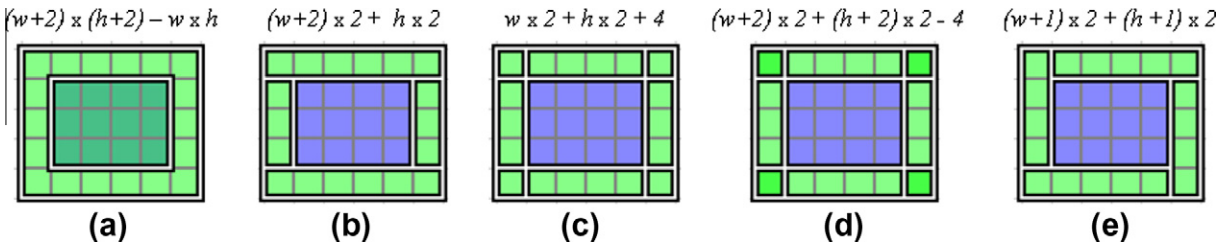

Fig. 2. 'Pond tiling' task constructions and associated rules: (a) the 'Area' strategy; (b) the 'I' strategy; (c) the '+4' strategy; (d) the '−4' strategy; (e) the 'Spiral' strategy.

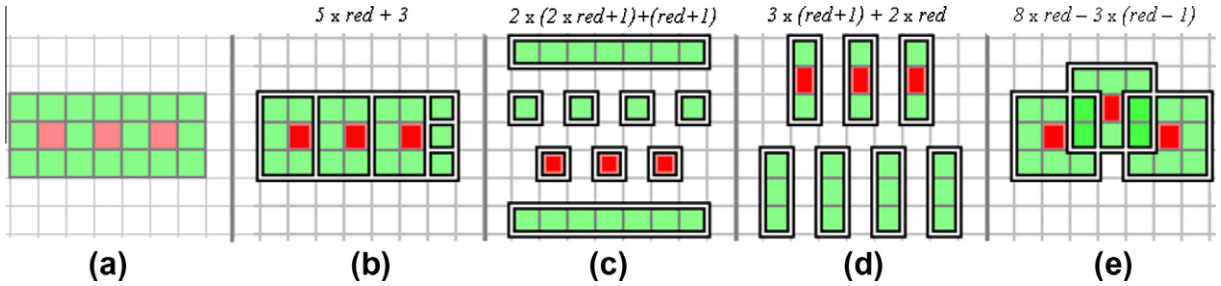

Fig. 3. 'Stepping Stones' task constructions and associated rules: (a) the task construction regardless of structure; (b) the 'C' strategy; (c) the 'HParallel' strategy; (d) the 'VParallel' strategy; (e) the 'Squares' strategy.

Most often learners start with specific constructions and gradually change their properties to make them general. The most important step is from a specific construction to making one of its components general, i.e. the transition from specific to partially general.

The use of property lists to construct patterns facilitates the derivation of the algebraic-like rule by the presence of the colouring property which refers to the number of tiles needed for certain parts of the construction; the rule is essentially formed by putting together the values of the colouring properties of all parts of a construction. Thus, the system supports simultaneously model construction and analysis.

To enable the dynamic presentation of a construction in the general world, the learners need to define a rule for the number of green tiles. This step was designed "to encourage students' reflection on their own actions. This process allows students to validate the generality of their final rule as well as a means to express their generalisations 'symbolically' " (Geraniou et al., 2009b, p. 55). Moreover, in collaborative activities the students are requested to share their constructions and rules and justify them to their peers. The aim of these activities is to emphasise the equivalence of seemingly different constructions and rules, which would deepen the learners' grasp of generalisation.

The construction in Fig. 1 and the rule in the bottom-right corner constitute one possible solution for the 'pond tiling' task. Although in its simplest form the rule is unique, there are several ways to build the construction and infer a rule from its components. Thus, there is no unique solution and students follow various kinds of strategies to build their constructions. More examples of such constructions and rules are displayed in Fig. 2. The 'Area' strategy in Fig. 2a is built by placing a larger rectangle over the pond; the '−4' strategy in Fig. 2d is built of rows and columns that correspond to the pond's width plus 2 and the ponds' height plus 2, respectively, and the highlighted corners (slightly darker) correspond to the overlapping between rows and columns.

We illustrate here another task called 'stepping stones', which is easier than the 'pond tiling' task, as it involves only one variable. It requires to build a construction such as the one in Fig. 3a and to find a rule for the green (lighter colour) tiles in relation to the red (darker) tiles,[1] i.e. the stepping stones. Some constructions are expanded for ease of visualisation and the variable "red" refers to the number of red tiles. In these figures, the internal structure of the constructions has been highlighted for clarity. In *eXpresser* all constructions would look the same in the normal course of the task.

As illustrated above, each task has multiple solutions corresponding to different visual representations. Some of these solutions are similar to each other while others are different. For example, in the 'pond-tiling' task the 'H' strategy in Fig. 1 and the '+4' strategy in Fig. 2c share similar characteristics because they have the same horizontal bars, while the 'I' strategy in Fig. 2b and '+4' strategy in Fig. 2c share similar characteristics for having the same vertical bars. The constructions are illustrated with same pond and the same 'stepping stones', respectively; however, learners build constructions of various dimensions. Therefore, in this work the notion of similarity between different strategies refers to structural similarity rather than the exact dimensions of the construction.

As mentioned above, the aim of collaborative activities is to reflect on the equivalence of seemingly different constructions and expressions. Fig. 4 illustrates two such constructions and expressions. The strategy used to solve the task is the same, but the visual representations of the construction and the representation of the variables involved in the task (the T-boxes) seem different. The learners' discussion aims to establish the equivalence of these

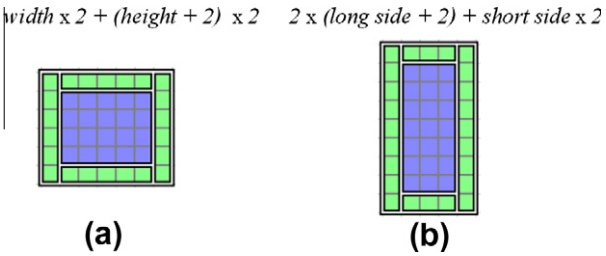

Fig. 4. Two approaches using the 'H' strategy.

[1] For interpretation of colour in Figs. 1–4, and 7, the reader is referred to the web version of this article.

representations by recognising what is different, e.g. the dimensions of the pond, the names used for variables, and what is the same and captures the essence of generalisation, e.g. the structure used (the 'H' strategy), the expressions. To this end, the detection of the strategies followed by each learner is necessary; this process is described in the next section.

4. Modelling user behaviour using a case-based representation

The architecture with the components involved in modelling the learners' behaviour when performing generalisation tasks and in the grouping mechanism are displayed in Fig. 5. As the learners are working with *eXpresser*, their actions are stored in individual Learner Models and passed to the Behaviour Analysis Module (BAM). Using the information from the Task Model, the BAM component analyses the actions of the learners and identifies the most similar approaches followed by learners when they do the task. This information is then passed back and stored in the Learner Models. To form groups for performing collaborative activities, the Grouping Module uses information about the approaches followed by learners, the so-called strategies as explained in Section 3, from the Task Model and about strategy or strategies used by each learner from the individual Learner Models.

Next we present a knowledge representation system to interpret relevant sequences of users actions during exploratory activities, and explain how inferences can be derived from these actions about ways users construct models/explore activities of mathematical generalisation. To this end, a case-based reasoning (CBR) approach is used; in CBR (Kolodner, 1993) knowledge is stored as cases, typically including the description of a problem and its solution. When a new problem is encountered, similar cases are retrieved and the solution is used or adapted from one or more of the most similar cases. The CBR cycle typically includes four processes (Kolodner, 1993): (a) *Retrieve* cases that are similar to the current problem; (b) *Reuse* the cases (and adapt) them in order to solve the current problem; (c) *Revise* the proposed solution if necessary; (d) *Retain* the new solution as part of a new case.

In exploratory learning the same problem can be solved through different user strategies and it is important to identify which strategy is used by the learner, as this relates to the learner's perception of generalisation tasks. To address this for *eXpresser* each task has a case-base (or knowledge base) of exploratory strategies. When a learner is performing an exploratory task, their construction-related actions are transformed into *a sequence of simple cases*, i.e. a strategy, and compared with all the strategies in the case-base for the particular task that the learner is working on; the case-base consists of general strategies, i.e. composite cases, rather than simple cases. To *retrieve* the strategies that are most similar to the one used by the learner, appropriate similarity metrics are employed that are described below. Once the most similar strategies are identified, they are used in the grouping mechanism that implements a form of *reuse* by taking this information into account along with the information on the strategies from the Task Model. The work described in this paper does not involve the *revise* and *retain* steps of the CBR cycle. These steps, however, are involved in identifying and storying new strategies in the Task Model. As this work is out of the scope of this paper, the reader is referred to (Cocea, Gutierrez-Santos, & Magoulas, 2009) for more details.

The following subsections present the knowledge representation and the similarity metrics used for strategy identification.

4.1. Case representation

In our approach, exploratory user strategies when performing a task are represented as a series of cases with certain relations between them.

Definition 1. A case is defined as $C_i = \{F_i, RA_i, RC_i\}$, where C_i represents the case and F_i is a set of attributes. RA_i is a set of relations between attributes and RC_i is a set of relations between C_i and other cases respectively.

Definition 2. The set of attributes of a given case C_i is defined as $F_i = \{\alpha_{i_1}, \alpha_{i_2}, \ldots, \alpha_{i_N}\}$.

It relates to user's construction and includes three types of attributes: (a) numeric, (b) variables and (c) binary. The numeric attributes correspond to the values in the property list and the variables correspond to the type of those properties: number, T-box, expression with number (s) or expression with T-box (es). The binary attributes refer to the membership of a case to a strategy and is defined as a *PartOfS* function which returns 1 if the case belongs to the strategy and 0 if it does not. There are S binary attributes, where S is the number of strategies in the knowledge base.

Definition 3. The set of relations between attributes of a given case C_i and attributes of other cases (as well as attributes of C_i) is represented as $RA_i = \{RA_{i_1}, RA_{i_2}, \ldots, RA_{i_M}\}$, where at least one of the attributes in each relation $RA_{i_m}, \forall m = \overline{1, M}$, is from F_i, the set of attributes of C_i.

Two types of binary relations are used: (a) *dependency relations* such as the one illustrated in Fig. 1 where the number of the iterations of the horizontal green patterns depends on the width of the pond (i.e. blue pattern) through the use of a T-box; (b) *value relations* such as the fact that the value of the colouring property of the horizontal green patterns in Fig. 1 is the value of the width of the pond. A case is considered *specific* when it does not have dependency relations and is considered *general* when it has all the dependency relations required by the task. The *specific* and *general* "status" of a construction is important in diagnosing the learner's stage in solving the task, as outlined in Section 3.

Definition 4. The set of relations between cases is represented as $RC_i = \{RC_{i_1}, RC_{i_2}, \ldots, RC_{i_P}\}$, where one of the cases in each relation $RC_{i_j}, \forall j = \overline{1, P}$ is the current case (C_i).

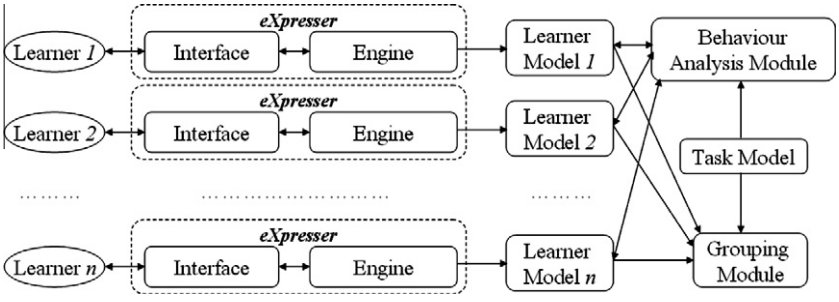

Fig. 5. Schematic architecture for *eXpresser* with the components involved in modelling the learners' strategies and in the grouping mechanism.

Two time-relations are used: (a) *Prev* relation indicates the previous case and (b) *Next* relation indicates the next case, with respect to the current case.

Definition 5. A user strategy is defined as $S_u = \{N_u(C), N_u(RA), N_u(RC)\}, u = \overline{1,r}$, where $N_u(C)$ is a set of cases, $N_u(RA)$ is a set of relation between attributes of cases and $N_u(RC)$ is a set of relations between cases.

4.2. Similarity assessment for case-based reasoning

The inference stage is based on scoring elements of the strategy followed by the learner to construct models in *eXpresser* with an emphasis on structural properties of the user's construction. To this end, the similarity between the current learner strategy and strategies previously stored is calculated taking into account their attributes and relations. The similarity metrics used for cases and strategies are displayed in Fig. 6. The overall similarity metric for strategies is: $Sim = w_1 * \overline{F_1} + w_2 * F_2 + w_3 * F_3 + w_4 * F_4$, where $\overline{F_1}$ is the normalised value of F_1. To bring F_1 in the same range as the other metrics, i.e. [0,1], we applied linear scaling to unit range (Aksoy & Haralick, 2001) using the function $\overline{F_1} = F_1/z$.

Weights are applied to the four metrics to express the central aspect of the construction, the structure. This is mostly reflected by the $\overline{F_1}$ metric and, to a lesser extent, by the F_3 metric. Therefore, the following values for the weights were set: $w_1 = 6$, $w_2 = 1$, $w_3 = 2$, $w_4 = 1$, leading to the range [0,10] for values of *Sim*.

The metrics have been tested for several situations of pedagogical importance: identifying complete strategies, mixed strategies, non-symmetrical strategies and partial strategies. The similarity metrics were successful in identifying all these situations – details can be found in (Cocea & Magoulas, 2009). For the purpose of this paper, we illustrate each of the situations mentioned above to facilitate a discussion about their potential role in group formation in Section 7.

Fig. 7 illustrates constructions for the four types of situations mentioned above. Fig. 7a illustrates a complete strategy; this is the 'H' strategy illustrated earlier in Fig. 1. Identifying complete strategies is important for two reasons: (a) to know if the learner has found a solution, and (b) to identify if that solution is specific, partially general or completely general, as a way to assess the learners' progress (see Section 3).

Fig. 7b illustrates a mixed strategy, in which parts of the 'I' and the '+4' strategies are used. From pedagogical point of view it is important to identify these situations so that guidance towards one of the strategies used can be given should the learners have difficulties to generalise.

Fig. 7c illustrates a non-symmetric strategy, which is, at the same time, a mixed strategy formed of parts of the 'H' and 'Spiral' strategies; the strategy in Fig. 7a is symmetric with respect to both horizontal and vertical axes, while the strategy in Fig. 7b is symmetric only with respect to the vertical axis. Symmetry facilitates generalisation, especially when learners need to 'translate' their construction into an algebraic-like expression. Therefore, detecting non-symmetric approaches is important so that guidance towards a symmetric approach can be given should the learners have difficulties to generalise.

Fig. 7d and e illustrate partial strategies, i.e. the pond is not completely surrounded; the strategy in Fig. 7e is also mixed, having parts from the 'H' and '+4' strategies. Detection of these situations is important for guiding learners by building on the strategy they started with should they be stuck or ask for help.

Similarity measures	Case	Strategy
Numeric attributes	$D_{IR} = \sqrt{\sum_{j=v+1}^{w} (\alpha_{I_j} - \alpha_{R_j})^2}$	$F_1 = \begin{cases} z / \sum_{i=1}^{z} D_{I_iR_i}, & \text{if } \sum_{i=1}^{z} D_{I_iR_i} \neq 0 \\ z, & \text{if } \sum_{i=1}^{z} D_{I_iR_i} = 0 \end{cases}$
Variables	$V_{IR} = \frac{\sum_{j=1}^{v} g(\alpha_{I_j}, \alpha_{R_j})}{v}$ $g(\alpha_{I_j}, \alpha_{R_j}) = \begin{cases} 1, & \text{if } \alpha_{I_j} = \alpha_{R_j} \\ 0, & \text{if } \alpha_{I_j} \neq \alpha_{R_j} \end{cases}$	$F_2 = (\sum_{i=1}^{z} V_{I_iR_i}) / z$
Relations between attributes	$P_{IR} = \frac{\lvert RA_I \cap RA_R \rvert}{\lvert RA_I \cup RA_R \rvert}$	$F_3 = (\sum_{i=1}^{z} P_{I_iR_i}) / y$
Relations between cases	$T_{IR} = \frac{\lvert RC_I \cap RC_R \rvert}{\lvert RC_I \cup RC_R \rvert}$	$F_4 = (\sum_{i=1}^{z} T_{I_iR_i}) / z$

I = Input Case/Strategy; R = Retrieved Case/Strategy
z = minimum number of cases of strategies I and R;
y = the number of cases in strategy R that have relations between attributes

Fig. 6. Similarity metrics for cases and strategies.

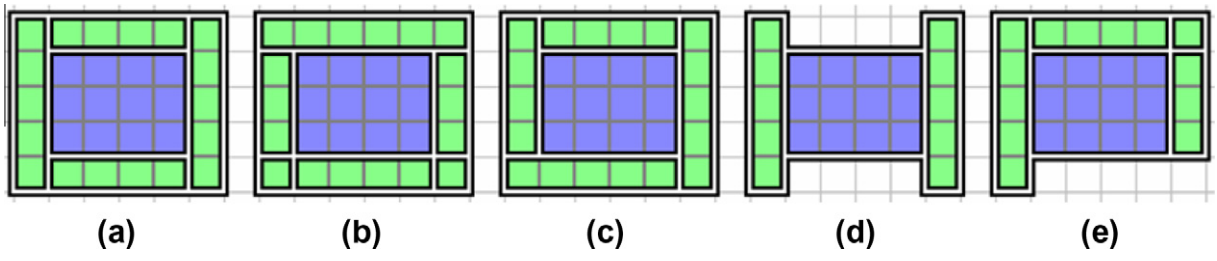

Fig. 7. Situations of pedagogical importance examples from the 'pond tiling' task: (a) complete strategy ('H' strategy); (b) mixed strategy ('I' and '+4'); (c) non-symmetric strategy; (d) partial strategy; (e) partial (mixed) strategy.

These different types of strategies could play an important role in creating collaboration learning activities for mathematical generalisation; a discussion on this aspect is presented in Section 7. The next section presents how user's strategy information and its similarity with other strategies is used for group formation.

5. Group formation for collaboration

In the context of *eXpresser*, the collaboration activities aim for students to reflect on the equivalence of seemingly different constructions and expressions and observe the similarity at a higher structural level as the essence of generalisation. These collaborative activities would benefit learners by raising their awareness of the several ways to approach the same task and their effort to establish the equivalence of representations for both visual patterns and variables expressed with T-boxes will lead to deeper understanding (Ainsworth, 1999; Petre et al., 1998; van der Meij & de Jong, 2006). Translating between representations, however, is found difficult by students (Ainsworth, 1999; Schoenfeld, Smith, & Arcavi, 2002; Yerushalmy, 1991). Therefore, our approach is to group students using the same or similar strategies, which ensures the presence of structural similarity between their constructions, to facilitate the translation between their representations of variables and expressions.

Unlike previous research, we propose characteristics of individual approaches and similarities between approaches as criteria for group formation, as these aspects are relevant for the way learning activities are defined in *eXpresser*. Although the goals of a task and user's performance or knowledge have been used in group formation, we are not aware of any research that considers users' behaviours, such as exploratory strategies, structural characteristics and relationships in users' constructions.

In order to incorporate our desired criteria in the grouping mechanism, we looked for a method that can perform grouping based on the defined criteria. Thus, we wanted to take into consideration the strategies used by the learners and the similarities between the various strategies used for the same task. Moreover, we are interested in a flexible way of defining similarities between strategies, as this has an impact on the difficulty of the collaborative task and also is viewed differently by different teachers (this is discussed in more detail in Section 7). Therefore, a mechanism that can handle this flexibility was required.

To this end, we are using clustering analysis approach that employs array-based clustering and resemblance coefficients. Similar approaches have been proposed in the literature for group formation in manufacturing systems (King & Nakornchai, 1982; Selim, Askin, & Vakharia, 1998; Joines, King, & Culbreth, 1996).

In array-based clustering, an incidence matrix is used whose entries are either zero or one. In our case, if the entry in row i and column j is one it means that learner j used strategy i; if it is zero, the learner has not used that strategy. The array-based technique leads to clusters of learners and strategies by rearranging the order of rows and columns to form diagonal blocks of ones in the incidence matrix. In our approach, similarity coefficients are used to form the incidence matrix and array-based clustering is then applied to obtain the clusters.

As mentioned at beginning of the section, it is important to group learners that use the same or similar strategies to facilitate the translation between different representations. To this end, the grouping procedure includes the following phases (see also Fig. 8):

Phase 1. Represent all strategies stored in the Task Model as binary vectors that define similarities between them.
Phase 2. Retrieve learner strategies from the Learner Models and represent learners as vectors whose elements depict the existence of a relation between a learner' strategy and a strategy stored in the set of task strategies.
Phase 3. Define resemblance coefficients and calculate them.
Phase 4. Derive the Strategies-Learners Matrix (SLM) from the results of previous step.
Phase 5. Perform clustering on SLM.

Definition 6. Let S be the set of strategies of a task: $S = \{s_j\}$, $j = 1, 2 \ldots, n$.

Every strategy can be represented as a n-dimensional vector of 0s and 1s: $s_j = \left(s_j^1, s_j^2, \ldots, s_j^n\right)$ where:

$$s_j^i = \begin{cases} 1 & \text{if } s_j \text{ is similar to strategy } s_i \\ 0 & \text{if } s_j \text{ is not similar to strategy } s_i \end{cases}$$

For example, the vectors for the five strategies of the 'pond-tiling' task illustrated in Fig. 2 are displayed in Table 1.

For example, as already mentioned in Section 2, strategy 'I' in Fig. 2b is similar to itself and to the '+4' strategy in Fig. 2c; the '+4' strategy is similar to itself and to strategies 'I' (Fig. 2b) and 'H' (Fig. 1). These similarities can be automatically deducted from the existence of structurally similar components like the ones illustrated in Section 2; alternatively they can be defined by teachers.

Definition 7. Let $L = \{\lambda_k\}$, $k = 1, 2, \ldots, m$ be the set of learners.

A learner can be represented as a vector of 0s and 1s: $\lambda_k = (\lambda_k^1, \lambda_k^2, \ldots, \lambda_k^n)$, where:

$$\lambda_k^i = \begin{cases} 1 & \text{if learner } \lambda_k \text{ used } s_i \text{ strategy} \\ 0 & \text{if learner } \lambda_k \text{ did not use } s_i \text{ strategy} \end{cases}$$

For example, learner A that has used the 'I' strategy is represented as (01000) and learner B that has used the 'Spiral' strategy is represented as (00001). Sometimes learners use combinations

Table 1
Vectors for the strategies of 'pond tiling' task.

	Area	'I'	'+4'	'−4'	'Spiral'
'Area'	1	0	0	0	0
'I'	0	1	1	1	0
'+4'	0	1	1	0	0
'−4'	0	1	0	1	0
'Spiral'	0	0	0	0	1

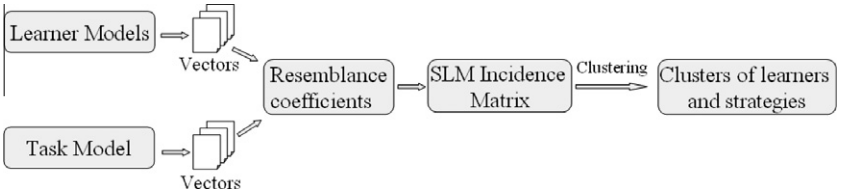

Fig. 8. The procedure for group formation.

of different strategies; for example, learner C who has used the 'I' and '+4' strategies would be represented as (01100). This vector formulation is based on the information stored in the Learner Models; thus, learner A and B have in their Learner Models that their constructions are most similar to strategies 'I' and 'Spiral', respectively, while the Learner Model for learner C indicates that the 'I' and '+4' strategies are most similar.

Definition 8. For each learner vector λ_k and each strategy vector s_j, the following are defined (see also Fig. 9):

1. a is the number of matching 1s, i.e. the number of strategies contained in both vectors;
2. b is the number of 1s in λ_k and 0s in s_j, i.e. the number of strategies followed by the learner which are contained in λ_k but not included in s_j;
3. c is the number of 0s in λ_k and 1s in s_j, i.e. the number of strategies that the learner did not follow but are included in s_j;
4. d is the number of matching 0s, i.e. the number of strategies that are not contained in both vectors.

5.1. Calculating resemblance coefficients between learners and their strategies

Two resemblance coefficients are used: one for the similarity between learners and strategies, and one for the relevance of each strategy for a particular learner.

Definition 9. The similarity coefficient (SC) between a learner λ_k and a strategy s_j is defined as: $SC(\lambda_k, s_j) = \frac{a}{a+b+c}$, for each learner $\lambda_k \in L$, $k = 1, 2, \ldots, m$ and each strategy $s_j \in S$, $j = 1, 2, \ldots, n$.

This was first defined for use in clustering by McAuley (1972) and is in fact a Jaccard similarity coefficient – a well known measure of similarity, which has been found to be quite robust, i.e. in several trials by Yin and Yasuda (2005) the results were within a small variation range around the average result, rather than within a wide range spanning from bad to very good results (that can still give a relatively good average).

Definition 10. The Relevance Coefficient (RC) of a strategy s_j for learner λ_k is defined as: $RC(\lambda_k, s_j) = \frac{a}{a+b}$, for each learner $\lambda_k \in L$, $k = 1, 2, \ldots, m$ and each strategy $s_j \in S$, $j = 1, 2, \ldots, n$.

The strategies-learners matrix is defined as:

$$SLM = \{c_{ij}\},\ i \in [1, n],\ j \in [1, m],$$

$$c_{ij} = \begin{cases} 1 & \text{if } RC \geqslant \theta^{RC} \quad \text{and} \quad SC \geqslant \theta^{SC} \\ 0 & \text{otherwise} \end{cases}$$

where $\theta^{RC}, \theta^{SC} \in (0, 1]$.

A minimum density of the matrix is necessary to obtain meaningful results. More specifically, each column should have at least a '1', i.e. each learner should follow at least one strategy. Therefore, the minimum density is the number of learners: m. Consequently, to fulfill the matrix density constraint, the values of θ^{RC} and θ^{SC} could be defined dynamically for each class. To avoid unnecessary computation, however, the following were established: (a) the value of θ^{RC} should not be lower than 0.5; this was decided because the relevance coefficient reflects the strategies followed by the learner and, consequently, should have an important role to ensure that the learner is placed in a group of learners that use at least one of the strategies followed by him/her; (b) calculate values dynamically only if the density constraint is not satisfied using the value of 0.5 for both thresholds. Therefore, the grouping starts with the value of 0.5 for both thresholds and if the matrix density constraint is not satisfied, the value of θ^{SC} is gradually decreased until the constraint is satisfied.

5.2. Array-based clustering

To illustrate the next phase of the procedure, i.e. the clustering, let us consider the matrix displayed in Step 1 of Fig. 10. Rank Order Clustering (ROC), one the most frequently used methods in array-based clustering (Joines et al., 1996), is applied, which involves organising columns and rows in the order of decreasing binary weights. The following procedure is applied which is illustrated in Fig. 10:

Step 1. Assign value 2^{m-j} to column j. Evaluate each row $\left(Row_i = \sum_{j=1}^{m} c_{ij} 2^{m-j}\right)$ and order rows in decreasing order. If there is no change compared to previous order, stop. Else, go to step 2.

Step 2. Assign value 2^{n-i} to row i. Evaluate each column $\left(Column_j = \sum_{i=1}^{n} c_{ij} 2^{n-i}\right)$ and order columns in decreasing order. If there is no change compared to previous order, stop. Else, go to step 1.

In this example, the following clusters were formed: learners 1, 3 and 6 with strategies 1 and 3; learners 2 and 5 with strategy 2, and learner 6 with strategy 4. In this particular example the blocks of 1s are clear cut; however, that is rarely the case, showing that clusters are not independent. Also, one strategy may be used by many learners, forming a big cluster. In the context of forming groups for collaboration using *eXpresser*, these situations are not critical limitations for the formulation of strategies-learners clusters: if clusters are not independent, it means that some learners are using other strategies besides the ones of that cluster; if many learners are using the same strategy, forming a large cluster, it can be broken down in several subgroups for the purpose of the collaborative task.

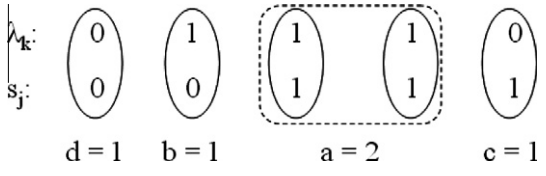

Fig. 9. Example for Definition 8.

Step 1

Strategies	Learners 1	2	3	4	5	6	
1	1	1	1	0	0	1	57
2	0	1	0	0	1	0	18
3	1	0	1	0	0	1	41
4	0	0	0	1	0	0	4
2^{m-j}	32	16	8	4	2	1	

Step 2

Strategies	Learners 1	2	3	4	5	6	2^{n-i}
1	1	1	1	0	0	1	8
3	1	0	1	0	0	1	4
2	0	1	0	0	1	0	2
4	0	0	0	1	0	0	1
	12	10	12	1	2	12	

Repeat Step 1

Strategies	Learners 1	3	6	2	5	4	
1	**1**	**1**	**1**	1	0	0	60
3	**1**	**1**	**1**	0	0	0	56
2	0	0	0	**1**	**1**	0	6
4	0	0	0	0	0	**1**	1
2^{m-j}	32	16	8	4	2	1	

Fig. 10. Steps of rank order clustering example.

The next section illustrates how the approach presented above has been used in conjunction with the strategy identification in a classroom application of *eXpresser*.

6. Classroom application

We illustrate the approach presented in the previous sections using data from a classroom session where 18 students used *eXpresser* to solve the 'stepping stones' task. Using the approach presented in Section 4, we identified that out of the 18 learners, 6 used the 'C' strategy (C) illustrated in Fig. 3b, 4 used the 'HParallel' strategy (H) displayed in Fig. 3c, 2 used the 'VParallel' strategy (V) illustrated in Fig. 3d, 1 used the 'Squares' strategy (S) displayed in Fig. 3d, 1 used a combination of 'HParallel' and 'VParallel' strategies (H&V) (Cocea & Magoulas, 2009) and the remaining 4 students were either off-task or used non-systematic approaches such as building the construction using individual tiles – see Table 2.

A subset of the vectors for strategies and learners is displayed in Table 3. For learners that used the same strategy, only one example is provided; for example, learners λ_1 to λ_6 have the same vectors and thus only learner λ_1 is displayed. The learners that did not follow a systematic approach are excluded. As shown in Table 2, four learners used non-systematic approaches to solve the task denoted by 'Other'. These could also be represented by a distinctive vector which will result in a cluster formed by these learners; however, they are already classified as a distinctive group and, therefore, including them in the grouping mechanism will only lead to unnecessary computations.

Table 4 displays the values of *RC* and *SC* for each strategy and learner. Using $\theta^{RC} = 0.5$ and $\theta^{SC} = 0.5$, the initial matrix in Table 5 is obtained; applying ROC to it leads to the final matrix in Table 5 and to the following groups:

(1) Group 1 includes learners λ_i, $i = 1, 2, \ldots, 6$ and λ_{13} that adopted the 'C' and 'Squares' strategies;
(2) Group 2 includes learners λ_i, $i = 7, 8, \ldots, 10$ and λ_{14} that adopted the 'HParallel' strategy;
(3) Group 3 includes learners λ_i, $i = 11, 12$ that adopted the 'VParallel' strategy.

The advantages of using this method, as opposed to clustering based only on the strategies used, is that the similarities between different strategies could be modified by the teacher. They could vary from being very strict (a strategy is similar only to itself) to being very relaxed (a strategy is similar to other strategies when there is at least one part that is similar). Given the way classes are formed in the UK, based on achievement levels, a relaxed definition of similarity would be more appropriate for higher achievements classes that need more challenges, while a strict definition of similarity would be more appropriate for lower achievement classes. Our proposed approach, thus, gives the necessary flexibility to teachers to define similarity depending on the characteristics of the class.

For example, a teacher may consider the 'Squares' strategy to be similar to the 'VParallel' rather than the 'C' strategy. Consequently, the strategies vectors would be: (a) 'C' strategy (1000); (b) 'HParallel' strategy (0100); (c) 'VParallel' strategy (0011); (d) 'Squares' strategy (0011). Using these vectors a new SLM matrix is obtained and the clustering procedure outputs the following groups:

(1) Group 1 includes learners λ_i, $i = 1, 2, \ldots, 6$ that used the 'C' strategy;
(2) Group 2 includes learners λ_i, $i = 7, 8, \ldots, 10$ and λ_{14} that used the 'HParallel' strategy;
(3) Group 3 includes learners λ_i, $i = 11, 12$ and λ_{14} that used the 'VParallel' and 'Squares' strategies.

Table 2
Distribution of strategies used by learners.

Strategies	C	H	V	S	H& V	Other
Number of learners	6	4	2	1	1	4

Table 3
Strategies and learners vectors.

Strategies	C	H	V	S	Learners	λ_1	λ_7	λ_{11}	λ_{13}	λ_{14}
C	1	0	0	1	C	1	0	0	0	0
H	0	1	0	0	H	0	1	0	0	1
V	1	0	0	0	V	0	0	1	0	1
S	1	0	0	1	S	0	0	0	1	0

Table 4
Similarity between strategies and learners and relevance of strategies for each learner.

Strategies		λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}	λ_{11}	λ_{12}	λ_{13}	λ_{14}
C forward	RC	1	1	1	1	1	1	0	0	0	0	0	0	1	0
	SC	0.5	0.5	0.5	0.5	0.5	0.5	0	0	0	0	0	0	0.5	0
HParallel	RC	0	0	0	0	0	0	1	1	1	1	0	0	0	0.5
	SC	0	0	0	0	0	0	1	1	1	1	0	0	0	0.5
VParallel	RC	0	0	0	0	0	0	0	0	0	0	1	1	0	0.5
	SC	0	0	0	0	0	0	0	0	0	0	1	1	0	0.5
Squares	RC	1	1	1	1	1	1	0	0	0	0	0	0	1	0
	SC	0.5	0.5	0.5	0.5	0.5	0.5	0	0	0	0	0	0	0.5	0

Table 5
Initial and final matrix. The bold values indicate the clusters.

Initial matrix															
Strategies	Learners														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
C	1	1	1	1	1	1	0	0	0	0	0	0	1	0
H	0	0	0	0	0	0	1	1	1	1	0	0	0	1
V	0	0	0	0	0	0	0	0	0	0	1	1	0	1
S	1	1	1	1	1	1	0	0	0	0	0	0	1	0

Final matrix (after ROC)														
Strategies	Learners													
	1	2	3	4	5	6	13	14	7	8	9	10	11	12
C	**1**	**1**	**1**	**1**	**1**	**1**	**1**	0	0	0	0	0	0	0
S	**1**	**1**	**1**	**1**	**1**	**1**	**1**	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	**1**	**1**	**1**	**1**	**1**	0	0
V	0	0	0	0	0	0	0	1	0	0	0	0	**1**	**1**

The mechanism we developed provides the teachers with groups based on the strategies followed by learners, i.e. the clusters formed as explained above. Using this information, teachers decide the size of groups and how the learners are distributed.

7. Discussion

The research presented in this paper had to address two main challenges: (1) assessment of individual behaviour in terms of recognising the strategies used and (2) developing a grouping mechanism that takes into account not only the information about the strategy or strategies followed by each learner, but also the similarity between the different strategies of a task.

The first challenge was addressed using a case-based reasoning approach that allowed the modelling of complex problems with multiple solutions. Unlike previous research that used concepts for modelling the domain and the student, we use *strategies* represented as series of cases with relations between them. This approach enabled a more detailed diagnosis of the learner as a strategy captures the complexity of an exploratory learning task more than domain concepts. Moreover, by including relations in the representations of the strategies, an essential part of the domain is included which cannot be captured by concepts alone.

As in the typical CBR new problems can be encountered for which there is no good match, in our context it is possible to encounter a strategy that is not part of the Task Model. There are two possibilities in this situation: (a) the new strategy is valid and should be stored in the Task Model, or (b) the new strategy is not valid and, therefore, should not be added to the knowledge base. The latter situation is not encountered in typical CBR and is only a characteristic of the domain we are working with. To address this, we have developed a mechanism that verifies new strategies that do not match the ones in the Task Model and stores them if the verification is successful – for details see (Cocea et al., 2009). Therefore, the presence of this mechanism enables to distinguish between learners that reached a valid solution despite this not being in the Task Model and learners that did not reach a solution or used non-systematic approaches, and ensures that the learners with valid new strategies are considered for the group formation.

One of the attributes of the simple cases refer to their membership to strategies. This attribute is useful in automatically defining similarities between strategies – if the criterion for this similarity is the existence of one or more common parts. The teachers, however, are given the option to override this definition and be able to set different similarities.

Regarding the second challenge, i.e. taking into account the similarities between strategies, as well as the strategies used by each learner in group formation, we used resemblance coefficients to define the similarity between learners and strategies and the relevance of each strategy for a particular learner, and an array-based clustering approach to form cluster of learners and strategies. The approach outputs homogeneous groups; however, heterogeneous groups can be formed by choosing one or more learners from each or some of the homogeneous groups.

The mechanism we propose has an important advantage compared with simple clustering, which is the flexibility given to teachers in defining similarities between strategies. When similarity is defined in a strict way, i.e. a strategy is similar only to itself, our mechanism gives the same output as a simple clustering method. The latter, however, does not allow a more relaxed definition of similarity, i.e. a strategy is similar to other strategies when there is at least one part that is similar or when there is some conceptual similarity, while our approach supports such definitions.

While the proposed clustering mechanism performs well, it has the limitation of using a 'black and white' approach rather than a continuous measurement scale. Thus, a strategy is either similar to another one or it is not similar at all. A grading scale, however, could be defined to reflect different degrees of similarity. In the same way, the similarity of a learner's strategy to all stored strategies is defined as either similar or dissimilar. It would be useful in the future to extend this mechanism to exploit information stored in the Learner Model about the most similar strategies, including similarity values for each one. For example, if a learner's strategy is similar to 'HParallel' and 'VParallel' strategies, with the values of 2.37 and 3.14, respectively, this information could be used instead of the 'black and white' approach.

Currently, our approach does not include social, cultural or personality factors, which are handled by the teacher. However, they could be easily integrated with our approach, which is part of our future work. Moreover, we will look into enabling teachers to set up constraints such as 'learner X should never be grouped with learner Y'.

The proposed approach does not distinguish between the types of strategies described in Section 4.2: complete, mixed, non-symmetric and partial, as we consider that mixing these types would be beneficial for learning. Consequently, the groups include learners using these different types. Often, the learners who follow complete strategies have used the other types before, and in practice teachers usually invite these learners to act as tutors for their peers. Research shows that peer tutors usually benefit by taking up this role because it helps them to reflect on their own knowledge and use it as a basis for constructing new knowledge – a process referred to as knowledge-building (Roscoe & Chi, 2007). Three properties of peer tutoring have been related to tutor learning: structuring, taking responsibility and reflecting (Biswas, Schwartz, Leelawong, & Vye, 2005). Giving explanations, asking and answering questions helps peer tutors in structuring their own knowledge; taking responsibility for their tutee's learning motivates peer tutors to gain a better understanding of the material; peer tutors' reflection on how their explanations were understood and used helps them in evaluating their own understanding of the domain.

Research also shows that tutee learning is maximised when the tutee reaches an impasse and is prompted to find the right way to continue and explain it, and is given an explanation only if they failed to do so (Vanlehn, Siler, Murray, Yamauchi, & Baggett, 2003). Therefore, learners with partial constructions that do not know how to continue would benefit from the explanations of a peer that has completed the same strategy; the ones with mixed strategies would benefit from discussing the similarities and differences between their approaches; the ones with non-symmetric strategies would learn about the benefit of working with a symmetric approach.

8. Conclusions

This paper presented an approach that extends the current research in the area of grouping for collaborative learning activities by replacing the use of a single indicator of performance with a more refined way of considering performance in the form of user strategies. This approach also complements the research that looks at the social characteristics of learners and the way they can be used in forming meaningful groups. Moreover, the social-based approaches can be integrated with our approach, potentially leading to a more comprehensive mechanism for group formation which we will investigate in the future.

Although we developed the approach for our particular exploratory learning environment, i.e. *eXpresser* and a specific domain, i.e. mathematical generalisation, the learner modelling mechanism and the grouping approach can be applied, either separately or in

combination, to other exploratory learning environments and domains characterised by multiple equally valid solutions.

Acknowledgments

This work was partially funded by the ESRC/EPSRC Teaching and Learning Research Programme (Technology Enhanced Learning; Award No: RES-139–25-0381).

References

Abbasi, A. A., & Younis, M. (2007). A survey on clustering algorithms for wireless sensor networks. *Computer Communications, 30*, 2826–2841. Network coverage and routing schemes for wireless sensor networks.

Ainsworth, S. (1999). The functions of multiple representations. *Computers & Education, 33*, 131–152.

Aksoy, S., & Haralick, R. (2001). Feature normalisation and likelihood-based similarity measures for image retrieval. *Pattern Recognition Letters, 22*, 563–582.

Aleven, V. (2003). Using background knowledge in case-based legal reasoning: A computational model and an intelligent learning environment. *Artificial Intelligence, 150*, 183–237.

Andaloro, G., & Bellomonte, L. (1998). Student knowledge and learning skill modeling in the learning environment 'forces'. *Computers & Education, 30*, 209–217.

Bednarz, N., Kieran, C., & Lee, L. (1991). Approaches to algebra: Perspectives for research and teaching. In N. Bednarz, C. Kieran, & L. Lee (Eds.), *Approaches to algebra: Perspectives for research and teaching* (pp. 3–12). Kluwer Academic Publishers.

Biswas, G., Schwartz, D., Leelawong, K., & Vye, N. (2005). Learning by teaching: A new agent paradigm for educational software. *Applied Artificial Intelligence, 19*, 363–392.

Bong, C.-W., & Rajeswari, M. (2011). Multi-objective nature-inspired clustering and classification techniques for image segmentation. *Applied Soft Computing, 11*, 3271–3282.

Brown, A., & Palincsar, A. (1989). Knowing, learning, and instruction. In L. Resnick (Ed.), *Guided cooperative learning and individual knowledge acquisition* (pp. 307–336). Hillsdale, NJ: Lawrence Erlbaum.

Bunt, A., & Conati, C. (2003). Probabilistic student modelling to improve exploratory behaviour. *User Modelling and User – Adaptive Interaction, 13*, 269–309.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*, 145–182.

Chun, S.-H., & Park, Y.-J. (2005). Dynamic adaptive ensemble case-based reasoning: Application to stock market prediction. *Expert Systems with Application, 28*, 435–443.

Cocea, M., & Magoulas, G. (2009). Task-oriented modeling of learner behaviour in exploratory learning for mathematical generalisation. In *Proceedings of the 2nd ISEE workshop* (pp. 16–24).

Cocea, M., Gutierrez-Santos, S., & Magoulas, G. (2009). Enhancing modelling of users' strategies in exploratory learning through case-base maintenance. In *Proceeding of the 14th UK workshop on case-based reasoning, UKCBR'09* (pp. 2–13).

Cohen, J. (1986). Theoretical considerations of peer tutoring. *Psychology in the Schools, 23*, 175–186.

Cohen, E. (1994). Restructuring the classroom: Conditions for productive small groups. *Review of Educational Research, 64*, 135.

Conati, C., & Merten, C. (2007). Eye-tracking for user modeling in exploratory learning environments: An empirical evaluation. *Knowledge-Based Systems, 20*, 557–574. Special Issue on Intelligent user interfaces.

Daradoumis, T., Guitert, M., Giménez, F., Marquès, J. M., & Lloret, T. (2002). Supporting the composition of effective virtual groups for collaborative learning. In *Proceedings of ICCE'02* (pp. 332–336). IEEE Computer Society.

Dhiraj, K., Rath, S., & Pandey, A. (2009). Gene expression analysis using clustering. In *3rd International conference on bioinformatics and biomedical engineering, 2009* (pp. 1–4). ICBBE, 2009.

Duke, R., & Graham, A. (2007). Inside the letter. *Mathematics Teaching Incoporating Micromath, 200*, 42–45.

Geraniou, E., Mavrikis, M., Hoyles, C., & Noss, R. (2008). A constructionist approach to mathematical generalisation. In M. Joubert (Ed.), *Proceedings of the British society for research into learning mathematics* (vol. 8). BSRLM proceedings.

Geraniou, E., Mavrikis, M., Kahn, K., Hoyles, C., & Noss, R. (2009b). Developing a microworld to support mathematical generalisation. In *PME 33: International group for the psychology of mathematics education* (vol. 29, pp. 49–56).

Geraniou, E., Mavrikis, M., Hoyles, C., & Noss, R. (2009a). Towards a constructionist approach to mathematical generalisation. *Research in Mathematics Education, 11*, 75–76.

Gogoulou, A., Gouli, E., Boas, G., Liakou, E., & Grigoriadou, M. (2007). Forming homogeneous, heterogeneous and mixed groups of learners. In *Proceedings of the workshop on personalisation in e-learning environments at individual and group level* (pp. 33 – 40).

Graf, S., & Bekele, R. (2006). Forming heterogeneous groups for intelligent collaborative learning systems with ant colony optimization. In M. Ikeda, K. D. Ashley, & T.-W. Chan (Eds.), *Intelligent tutoring systems. LNCS* (vol. 4053, pp. 217–226). Springer.

Gregory, R. P. (1984). Streaming, setting and mixed ability grouping in primary and secondary schools: Some research findings. *Educational Studies, 10*, 209–226.

Grigoriadou, M., Samarakou, M., Mitropoulos, D., Rigoutsos, A., Stavridou, E., & Solomonidou, C. (1999). Vectors in physics and mathematics. In *Proceedings of the international conference on technology and education (ICTE)* (pp. 71–73). Edinburgh.

Han, S.-G., Lee, S.-G., & Jo, S. (2005). Case-based tutoring systems for procedural problem solving on the WWW. *Expert Systems with Applications, 29*, 573–582.

Hoyles, C., & Küchemann, D. (2002). Students understanding of logical implication. *Educational Studies in Mathematics, 51*, 193–223.

Huang, M.-J., Huang, H.-S., & Chen, M.-Y. (2007). Constructing a personalized e-learning system based on genetic algorithm and case-based reasoning approach. *Expert Systems with Applications, 33*, 551–564.

Johnson, D., & Johnson, R. (1994). *Learning together and alone: Cooperative, competitive, individualistic learning*. Allyn & Bacon, Boston, MA.

Joines, J. A., King, R. E., & Culbreth, C. T. (1996). A comprehensive review of production-oriented manufacturing cell formation techniques. *International Journal of Flexible Automation and Integrated Manufacturing, 3*, 225–265.

King, J., & Nakornchai, V. (1982). Machine-component group formation in group technology: Review and extension. *International Journal of Production Research, 20*, 117–133.

Koedinger, K., Anderson, J.R., Hadley, W.H., & Mark, M. (1997). *Intelligent tutoring goes to school in the big city*. International Journal of Artificial Intelligence in Education, 8, 30–43.

Kolodner, J. (1993). *Case-based reasoning*. Morgan Kaufmann Publishers, Inc..

Küchemann, D. (1991). Algebra. In K. Hart (Ed.), *Childrens understanding of mathematics: 11–16* (pp. 102–119). London: John Murray.

Kumar, P., Gopalan, S., & Sridhar, V. (2005). Context enabled multi-CBR based recommendation engine for e-commerce. In *Proceedings of the IEEE international conference on e-business engineering (ICEBE)* (pp. 237–244). IEEE Press.

Leornard, J. (2001). How group composition influenced the achievement of sixth-grade mathematics students. *Mathematical Thinking and Learning, 3*, 175–200.

Lynch, C., Ashley, K., Aleven, V., & Pinkwart, N. (2006). Defining "ill-defined domains"; A literature survey. In *Proceedings of the workshop on intelligent tutoring systems for ill-defined domains at the 8th ITS conference* (pp. 1–10).

Macintyre, H., & Ireson, J. (2002). Within-class ability grouping: Placement of pupils in groups and selfconcept. *British Educational Research Journal, 28*, 249–263.

Malara, N., & Navarra, G. (2003). *ArAl Project: Arithmetic pathways towards favouring pre-algebraic thinking*. Pitagora Editrice, Bologna.

Mason, J. (2002). Generalisation and algebra: Exploiting children's powers. In L. Haggarty (Ed.), *Aspects of teaching second ary mathematics: Perspectives on practice* (pp. 105–120). Routledge Falmer and the Open University.

Mastropieri, M. A., Scruggs, T. E., & Graetz, J. E. (2003). Reading com prehension instruction for secondary students: Challenges for struggling students and teachers. *Learning Disability Quarterly, 26*, 103–116.

Matsuda, N., & VanLehn, K. (2005). Advanced geometry tutor: An intelligent tutor that teaches proof-writing with construction. In *Proceedings of the 12th international conference on artificial intelligence in education 443450*. IOS Press, Amsterdam.

McAuley, J. (1972). Machine grouping for efficient production. *Production Engineer, 51*, 53–57.

Mcmaster, K., Fuchs, D., & Fuchs, L. (2006). Research on peer-assisted learning strategies: The promise and limitations of peer-mediated instruction. *Reading and Writing Quarterly, 22*, 5–25.

Melis, E., Büdenbender, J., Goguadze, G., Libbrecht, P., & Ullrich, C. (2003). Knowledge representation and management in activemath. *Annals of Mathematics and Artificial Intelligence, 38*, 47–64.

Merten, C., & Conati, C. (2006). Eye-tracking to model and adapt to user meta-cognition in intelligent learning environments. In *IUI '06: Proceedings of the 11th international conference on intelligent user interfaces* (pp. 39–46). New York, NY, USA: AMC.

Moss, J., & Beatty, R. (2006). Knowledge building in mathematics: Supporting collaborative learning in pattern problems. *International Journal of Computer-Supported Collaborative Learning, 1*, 441–465.

Noss, R., Healy, L., & Hoyles, C. (1997). The construction of mathematical meanings: Connecting the visual with the symbolic. *Educational Studies in Mathematics, 33*, 203–233.

Noss, R., & Hoyles, C. (1996). *Windows on mathematical meanings: Learning cultures and computers*. Kluwer Academic Publishers.

Noss, R., Hoyles, C., Mavrikis, M., Geraniou, E., Gutierrez-Santos, S., & Pearce, D. (2009). Broadening the sense of 'dynamic: A microworld to support students mathematical generalisation. *Special Issue of the Int J Math Educ (ZDM): Transforming Mathematics Education through the Use of Dynamic Mathematics Technologies, 41*, 493–503.

Petre, M., Blackwell, A. F., & Green, T. R. G. (1998). Cognitive questions in software visualization. In J. Stasko, J. Domingue, M. Brown, & B. Price (Eds.), *Software visualization: Programming as a multi-media experience* (pp. 453–480). MIT Press.

Robinson, D., Schofield, J., & Steers-Wentzell1, K. (2005). Peer and cross-age tutoring in math: Outcomes and their design implications. *Educational Psychology Review, 17*, 327–362. 10.1007/s10648-005-8137-2.

Rohrbeck, C. A., Ginsburg-Block, M. D., Fantuzzo, J. W., & Miller, T. R. (2003). Peer-assisted learning interventions with elementary school students: A meta-analytic review. *Journal of Educational Psychology, 95*, 240–257.

Roscoe, R. D., & Chi, M. (2007). Understanding tutor learning: Knowledge-building and knowledge-telling in peer tutors explanations and questions. *Review of Educational Research, 77*, 534–574.

Ross, A. J., Tojeiro, R., & Percival, W. J. (2011). Understanding the faint red galaxy population using large-scale clustering measurements from SDSS DR7. *Monthly Notices of the Royal Astronomical Society, 413*, 2078–2086.

Schoenfeld, A. H., Smith, J. P., & Arcavi, A. (2002). Student perceptions of aspects of algebraic function using multiple representation software. In R. Glaser (Ed.). *Advances in instructional psychology* (vol. 4, pp. 55–175). Hillside, NJ: Erlbaum.

Selim, H. M., Askin, R. G., & Vakharia, A. J. (1998). Cell formation in group technology: Review, evaluation and directions for future research. *Computers & Industrial Engineering., 34*, 3–20.

Slavin, R. (1995). *Cooperative learning*. Needham Heights, MA: Allyn & Bacon.

Slavin, R.E. (2003). When and why does cooperative learning increase achievement? In H. Daniels, & A. Edwards (Eds.), *The RoutledgeFalmer Reader in psychology of education. RoutledgeFalmer Readers in education* (vol. 1, pp. 271–293). RoutledgeFalmer.

Stathacopoulou, R., Magoulas, G. D., Grigoriadou, M., & Samarakou, M. (2005). Neuro-fuzzy knowledge processing in intelligent learning environments for improved student diagnosis. *Information Sciences, 170*, 273–307.

Stottler, R., & Ramachandran, S. (1999). A case-based reasoning approach to internet intelligent tutoring systems (ITS) and ITS authoring. In *Proceedings of the twelfth international Florida artificial intelligence research society conference* (pp. 181–186). AAAI Press.

Ting, C.-Y., & Phon-Amnuaisuk, S. (2009). Factors influencing the performance of dynamic decision network for INQPRO. *Computers & Education, 52*, 762–780.

Ushioda, A., & Kawasaki, J. (1996). Hierarchical clustering of words and application to NLP tasks. In E. Ejerhed, & I. Dagan (Eds.), *Fourth workshop on very large corpora. Association for computational linguistics, Somerset, New Jersey* (p. 2841).

van der Meij, J., & de Jong, T. (2006). Progression in multiple representations: Supporting students' learning with multiple representations in a dynamic simulation-based learning environment. In *The EARLI SIG text and graphics comprehension bi-annual meeting 2006, EARLI SIG2*.

van Joolingen, W. R., King, S., & de Jong, T. (1997). The simquest authoring system for simulation-based discovery environments. In B. de Boulay & R. Mizoguchi (Eds.), *Knowledge and media in learning systems* (pp. 79–87). Amsterdam: IOS.

Vanlehn, K., Siler, S., Murray, C., Yamauchi, T., & Baggett, W. B. (2003). Why do only some events cause learning during human tutoring. *Cognition and Instruction, 21*, 209–249.

Veermans, K.H. (2003). *Intelligent support for discovery learning*. PhD thesis, University of Twente.

Vermette, P. (1998). *Making cooperative learning work: Student teams in K-12 classrooms*. Merrill, Upper Saddle River, NJ.

Warren, E., & Cooper, T. J. (2008). Generalising the pattern rule for visual growth patterns: actions that support 8 year olds' thinking. *Educational Studies in Mathematics, 67*, 171–185.

Webb, N. M., Baxter, G., & Thompson, L. (1997). Teachers' grouping practices in fifth-grade science classrooms. *Elementary School Journal, 98*, 91–124.

Yerushalmy, M. (1991). Student perceptions of aspects of algebraic function using multiple representation software. *Journal of Computer Assisted Learning, 7*, 42–57.

Yin, Y., & Yasuda, K. (2005). Similarity coefficient methods applied to the cell formation problem: a comparative investigation. *Computers & Industrial Engineering, 48*, 471–489.

Zurita, G., Nussbaum, M., & Salinas, R. (2005). Dynamic grouping in collaborative learning supported by wireless handhelds. *Educational Technology and Society, 8*, 149–161.