Integrating Knowledge Gained From Data Mining With Pedagogical Knowledge

Roland Hübscher[1] and Sadhana Puntambekar[2]
rhubscher@bentley.edu, puntambekar@education.wisc.edu
[1] Department of Information Design and Corporate Communication, Bentley College
[2] Department of Educational Psychology, University of Wisconsin

Abstract. Discovering knowledge from raw data is one of the goals of data mining. Yet, it is not always clear how this knowledge is used in educational computing systems and how exactly it is integrated with other knowledge like the pedagogy used. We present a case study where the use of the data mining results was initially described, to a large degree, at the implementation level, thus largely ignoring the nature of the different kinds of knowledge involved. Based on Clancey's heuristic classification model [7], the description is raised to a conceptual level, the knowledge level. This results in an explicit and well-defined integration of knowledge discovered with data mining techniques, pedagogical knowledge and linguistic knowledge. Such a knowledge-level description leads to an improved understanding of the system and its effects on the learners.

1 Introduction

Educational data are mined with the goal to discover knowledge about the learners, educational software and other classroom interventions. Thus, the designers need to be explicit about how that knowledge is being used to redesign educational software. Yet, many of us working in the general area of educational technology too often talk about software or more general interventions at the implementational level. Staying at that level leaves the use of the data mining knowledge and its integration with pedagogical knowledge implicit.

Having struggled with this issue ourselves, we present in this paper a case study from our own research. CoMPASS is an educational hypertext system helping middle-school students learn science. Originally, we had used data mining to better understand the design of CoMPASS, most notably the impact of a concept map as navigation support [16, 18]. Analysis using data mining techniques showed that certain navigation patterns are indicative of student learning. We then decided to take advantage of these patterns and use them to provide the students with adaptive and context sensitive prompts which are consistent with the notion of scaffolding [17]. Prompting a student with a question or a statement is a tool frequently used by teachers to make the student reflect on issues that will hopefully result in an improved understanding or problem solving behavior. Based on the user categorization found with data mining, we developed a simple scheme to generate such prompts [16]. However, the mechanism for generating the prompt was quite ad hoc and, with hindsight, it was not clear enough what roles pedagogy and insights gained from data mining played. This was one of the main reason its implementation stalled. Therefore, we developed a framework for our prompt-generating scheme addressing these issues.

It must always be clear what knowledge led to which design decisions. For instance, was there a valid pedagogical reason to make the user select between two choices, or was it computationally infeasible to reduce the "choice" to one item? Thus, design decisions and their reasons have to be made explicit. Designers of adaptive hypermedia systems [4], for instance, have dealt with this issue. Should an educational adaptive hypermedia system have an explicit pedagogical model describing pedagogical knowledge, or is it fine to fold the pedagogical assumptions into the adaptive engine itself? Quite a lively discussion around this topic happened in an after-session meeting at Adaptive Hypermedia 2004. The preferred, though surprisingly not unanimous, view was that indeed, the pedagogical model should be separate. Another example is the confusion between the concept of prerequisite and pedagogy [11]. The fact that a concept *A* is a prerequisite for understanding a concept *B* does not imply that *A* must be taught before *B* as some adaptive hypermedia systems assume [5]. Again, different kinds of knowledge are mixed up which leads to important implicit design decisions that should have been made explicit.

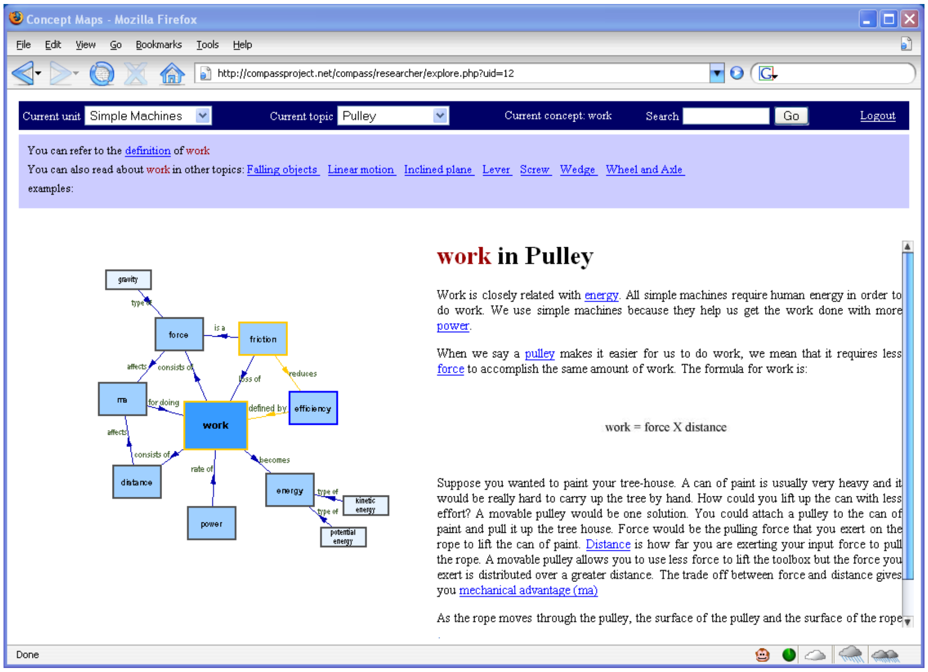
In this paper, we will present the case study as follows. First, we will introduce the educational hypertext system CoMPASS and the design of scaffolding with adaptive prompts. Therefore, scaffolding in educational systems will be briefly described before presenting the original design described at the implementational level. We then show how this description can be raised to the conceptual knowledge level making explicit the role the different kinds of knowledge play and how they can be integrated. Although this is a specific case study, it will become quite apparent how other designs suffering from similar problems can take advantage of the approach described here.

2 Scaffolding in CoMPASS with Prompts

2.1 CoMPASS

CoMPASS, shown in Figure 1, is a hypertext system to help middle-school students learn science. CoMPASS presents students with two representations, a textual representation of concept descriptions and a graphical representation in the form of a navigable concept map. Both representations change dynamically as students traverse through the domain and make reading choices.

Each page in CoMPASS is a description of a concept within a certain topic. In a domain like physics such concepts would include force, mass, acceleration, etc. When students choose a concept, CoMPASS presents them with a description of that concept along with a navigable map that shows them the related concepts. Instead of presenting a hierarchically organized set of topics with the option of going to a glossary page for related information, the very basis for navigation in CoMPASS is associative relationships between conceptual units. The maps provide local coherence, by showing students the concepts that are most related to the concept that they have chosen. In CoMPASS, the navigable maps are drawn with a fisheye, with the concept that the student has chosen as the focus. The rest of the map is drawn by retrieving the most related concepts from a database, which stores the semantic relatedness information of the concepts. We have used the relationship strength to



You can refer to the definition of work

You can also read about work in other topics: Falling objects Linear motion Inclined plane Lever Screw Wedge Wheel and Axle

examples:

work in Pulley

Work is closely related with energy. All simple machines require human energy in order to do work. We use simple machines because they help us get the work done with more power.

When we say a pulley makes it easier for us to do work, we mean that it requires less force to accomplish the same amount of work. The formula for work is:

work = force X distance

Suppose you wanted to paint your tree-house. A can of paint is usually very heavy and it would be really hard to carry up the tree by hand. How could you lift up the can with less effort? A movable pulley would be one solution. You could attach a pulley to the can of paint and pull it up the tree house. Force would be the pulling force that you exert on the rope to lift the can of paint. Distance is how far you are exerting your input force to pull the rope. A movable pulley allows you to use less force to lift the toolbox but the force you exert is distributed over a greater distance. The trade off between force and distance gives you mechanical advantage (ma)

As the rope moves through the pulley, the surface of the pulley and the surface of the rope

Figure 1. CoMPASS with the concept map supporting navigation on the left and the content (concept ‘work’ as it is used in the context of the simple machine‘ pulley’) on the right.

determine the spatial proximity of the concepts. Thus the stronger the relationship between the two concepts, the closer they are spatially in the concept map.

In CoMPASS, students can easily switch views to go to a related topic. This provides global coherence, because students can see what other topics they can go to that could be related to a particular topic. In addition, they can also view a particular concept from multiple perspectives as described below. For example, a student setting up an experiment with a pulley might be interested in learning about ’work’ in the context of a lever instead of pulley as shown in Figure 1. Thus, the student can navigate within a context (e.g., pulley) or across (e.g., from pulley to lever). Learning in a subject area, such as science, involves understanding the rich set of relationships among important concepts, which may form a web or a network. Revisiting the same material at different times, in rearranged contexts, for different purposes, and from different conceptual perspectives is essential for attaining the goals of advanced knowledge acquisition [22]. The alternative views that CoMPASS offers can help students to study science concepts and phenomena in depth by visiting them in multiple contexts.

3 Scaffolding with Prompts

A next step is to add adaptive support with textual prompts that help students directly, especially when they have some problems, and indirectly also teachers who cannot attend to all students in the classroom at all times. The prompts are supposed to *scaffold* the students. Since the concept of scaffolding has been somewhat overused [17], we briefly describe it.

Scaffolding in the context of learning has originally been defined as an "adult controlling those elements of the task that are essentially beyond the learner's capacity, thus permitting him to concentrate upon and complete only those elements that are within his range of competence" [27]. Scaffolding has been linked to the work of Soviet psychologist Lev Vygotsky, although he never used the term scaffolding. According to Vygotsky, a novice learns with an expert, and learning occurs within the novice's Zone of Proximal Development (ZPD). ZPD is defined as the "distance between the child's actual developmental level as determined by independent problem solving and the higher level of potential development as determined through problem solving under adult guidance and in collaboration with more capable peers" [25]. Enabling the learner to bridge this gap between the actual and the potential depends on the resources or the kind of support that is provided. Instruction in the ZPD can therefore be viewed as taking the form of providing assistance or scaffolding, enabling a child or a novice to solve a problem, carry out a task or achieve a goal "which would be beyond his unassisted efforts" [27].

Proper scaffolding requires a computer-based learning environment like CoMPASS to support, among other things, (a) continuous assessment of the learner needs to be used to calibrate the support; (b) scaffolding fading away over time and the learner taking control of the task; and (c), the learner needing to be actively involved in the learning process [15,23].

The implications for the prompts are therefore: (a) the prompts must be adapted to the student's current understanding and progress, i.e., they must be adaptive, context sensitive and individualized; (b) the prompts should be formulated and presented (or not!) so that the student is not "bothered" by them when there is no need for support anymore; and (c), the prompts should be formulated such that they result in active reflection and they are not just corrective suggestions to be followed mindlessly.

Adaptive support requires modeling users as in, for instance, adaptive hypermedia systems where mostly explicit user models are used by the system to adapt presentation and navigation support to each individual user [4]. However, this is simply not feasible given how CoMPASS is being used. Only sparse user data is available and there is no time to collect detailed user information with questionnaires or multiple-choice tests. We basically have to rely on a few clicks to detect how a student is progressing.

Fortunately, earlier work on data mining the navigation data collected from the CoMPASS users had revealed that students using CoMPASS can be assigned to categories that can be associated with the students' approach to learning and understanding of the material [19].

Since this paper's focus is on integration of the discovered knowledge and not on the discovery process itself, we just briefly summarize the data mining methods used. To find the learner categories, each student's clickstream was converted into a navigation matrix N describing the number of transitions $N_{i,j}$ from concept c_i to concept c_j and then pruned using the Pathfinder algorithm [21]. The resulting matrices, one for each user, were then clustered using the k-Means algorithm [10]. The students in these clusters were then analyzed to see what educational characteristics they had in common. For instance, as described in [19], the students in one cluster showed that they focused on the relevant (as determined by an expert) concepts but also visited related concepts. Such students tended to do well. In another cluster, students apparently had no well-defined focus and explored concepts also in other topics. Yet another group showed a random behavior indicating that

Condition: The ratio of visited to existing related topics for a concept *C* is small.
Justification: The student does not visit a certain concept *C* in related topics. Encourage student to read the same concepts in related topics.
Example prompt: Have you read concept *C* in any other topic?

Figure 2. A rule described at the implementational level.

they were not aware of the conceptual structure of the domain.

Based on these results, we developed rules to generate adaptive prompts [16]. These rules use some simple characteristics of the realtime navigation data, i.e., the clickstream, to detect behavior associated with the clusters found during data mining. An example of such a rule is shown in Figure 2. Of course, this is not the implementation of the rule itself, but it is described in terms of concepts at the level of the implementation, especially the condition. The justification is not used by the system and only serves as a comment to the writer of the rule.

This approach sounds reasonable and is described in more general form in Figure 3. However, as the figure shows, the role that the pedagogical knowledge plays and how it connected to the classifications generated by the data mining methods is conceptually not very clear. The reason is that the approach is described at the implementational level instead of the knowledge level. Thus, modifications to the data mining approach, to the pedagogical approach and to what kinds of text prompts to use will have to happen at the implementational level.

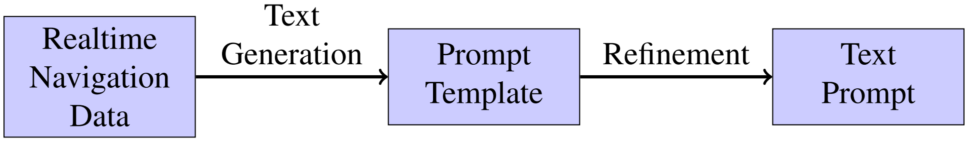


Figure 3. Text prompt generation algorithm.

4 Integrating Data Mining with Pedagogical Knowledge

One of the pitfalls of system design is to make decisions at the implementational level when they should be made at the conceptual level, although, of course, implementational constraints need to be considered. For instance, in the early AI days, expert system developers argued about forward versus backward chaining, instead of focussing on what tasks the experts solved and with what problem solving methods. Fortunately, the discussion soon moved from the implementational level to the conceptual, or, knowledge level [26]. This type of work benefited from earlier research by Clancey on classification of the reasoning that goes on in expert systems [7].

Clancey [7] describes the simple "heuristic classification" inference structure in Figure 4 that provides the basis for many problem solving methods used by experts. For instance, from a set of symptoms (data) a doctor abstracts to a class of symptoms (data

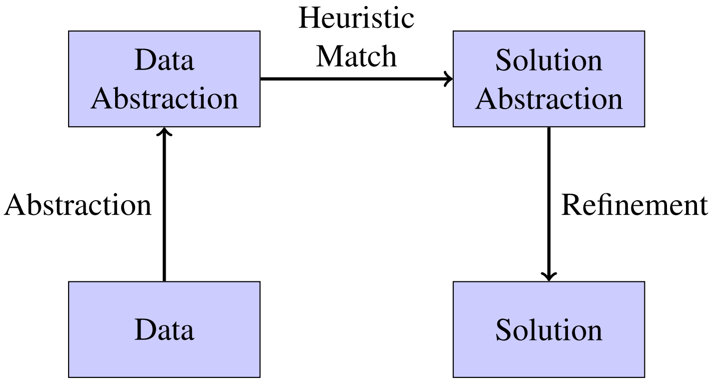


Figure 4. Heuristic classification.

abstraction) which then requires a certain type of treatment (solution abstraction) which is found by applying medical knowledge (heuristic match). Using contextual information about the patient and treatment, the type of treatment can be refined, e.g., the dosage can be adjusted (refinement).

More direct and simpler symptom-to-treatment reasoning is not as good, since it does not take advantage of the different types of knowledge (classification, refinement, heuristic medical knowledge) and it would result in a much less effective representation of the knowledge. Adding new knowledge would be messy since it would not be clear how exactly it should be integrated and used. Just reacting to symptoms before one is able to classify the type of a problem would also result in more mistreatment. If new medications or symptom detectors are introduced, it is clear how they are going to be integrated in the heuristic classification scheme.

Figure 5 shows the heuristic classification scheme applied to the generation of adaptive prompts. The direct link from the realtime navigation data to the text prompts (see Figure 3) has been replaced by a conceptual structure integrating the various types of knowledge involved. It should not be a surprise that the heuristic classification scheme is such a good match. After all, the learner's navigation behavior is the symptom and the text prompt the treatment for the learner.

In Figure 5, there are three types of knowledge explicitly represented. The knowledge discovered with the data mining methods is represented by the learner categories. Pedagogical knowledge is used to decide what kind of textual interventions should be used. And finally, linguistic knowledge is used to create the appropriate text. The latter is especially important if several interventions need to be combined into one phrase, or past prompts need to be taken into account.

Applying the problem solving method in Figure 5 results in the rule set shown in Figure 6. There are now three steps instead of one. First, the student's navigation behavior is still categorized the same way as in the original version, but this time mapped explicitly to one or more of the learner categories discovered via data mining of the off-line data collected in the past. Since the categories are the result of the data mining process, once improved data mining results are available—and we are working on them—these categories

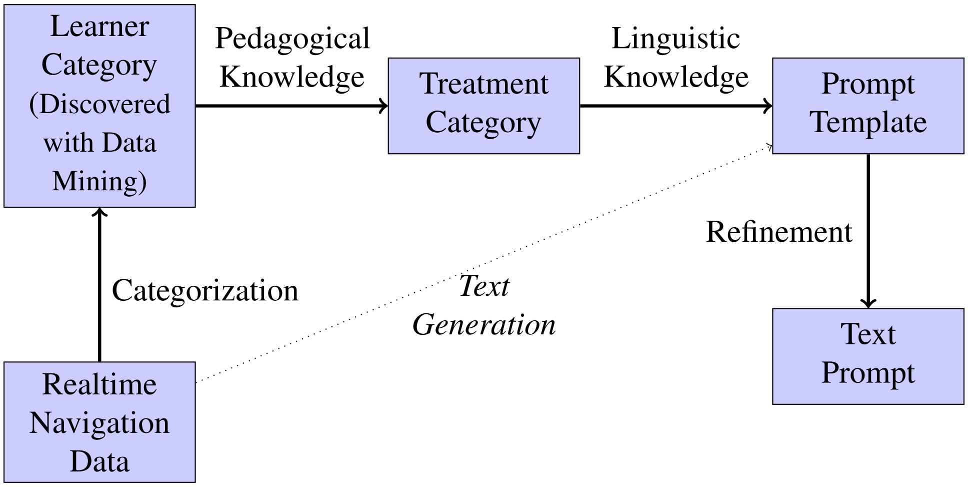


Figure 5. Text prompt generation at the conceptual level using the heuristic classification scheme. The dotted line refers to Figure 3.

can be modified.

Second, based on the learner categories and the used pedagogy, the kind of feedback deemed most useful for such a learner is suggested. Although we intend to use as simple rules as shown in Figure 6, a much more complicated reasoning process could be involved, however, using pedagogical knowledge only. Third, the treatments are collected and translated into a proper prompt. For instance, if the first rule (categorization) in Figure 6 applies to two concepts, the second linguistic rule will be applied. We will use a relatively simple yet quite powerful template-based approach to generate the natural language output [8,20]. This will also allow us to take previous prompts into consideration and avoid repeating the same prompt over and over even if the student does not improve. In our approach, Clancey's heuristic match (see Figure 4) is composed of two steps. The refinement step is the same as in the old case where the necessary variables in the template are bound based on the context.

In the original formulation of how the prompts were generated, all three kinds of knowledge were mixed into each rule. Theoretically as well as practically, this is a problem. From a theoretical point of view, it is quite unclear what types of knowledge are involved and how. For instance, the importance of the linguistic knowledge was originally overlooked. From a practical point of view, the modular approach makes it clear what to change, be the change either due to modification in the pedagogical approach or due to improved data mining or text generation methods.

This approach is not only suitable for the specific situation in CoMPASS. An obvious place to apply the approach described in this paper is in the context of association rules [9, 12]. These rules capture associations like *beer* ⇒ *diapers* between variables of items in shopping baskets [2]. Based on our experience in CoMPASS, it will be useful to always carefully consider and make explicit what type of knowledge such rules capture. Looking at the current literature, it seems, that this is generally not done in an explicit and rigorous way. Such association rules are normally judged by some mathematical definition

Categorization
Condition: The ratio of visited to existing related topics for a concept C is small
Action: Assign student to category `ignores_related_topics`

Pedagogical knowledge (heuristic match)
Condition: Student `ignores_related_topics`
Action: Say something to encourage reading concept C in other topics.

Linguistic knowledge (heuristic match)
Condition: Encourage reading concept C in other topics
Action: Create prompt "Have you read concept C in any other topic?"

Condition: Encourage reading concept C_1 in other topics *and* encourage reading concept C_2 in other topics
Action: Create prompt "Have you read concepts C_1 and C_2 in any other topic?"

Figure 6. A set of rules based on the conceptual description shown in Figure 5.

of interestingness [1, 3, 14] which is perfectly fine for finding the rules. However, the rules also need to be used in a meaningful way and that is where our approach may be useful.

5 Conclusions

We have presented a case study of our own research showing that making the different kinds of knowledge including the ones gained from data mining can have great benefits. It is clear what modifications need to be made if pedagogy, data mining techniques, or other parts are being changed. This is especially useful if weak points need to found in the educational software. It may not be *what* the prompts say, but *how* they say it. This is similar to the difficulty students have with word problems where often difficulties reflect students' language problems but not necessarily their math problems [24].

Although this case study presented a specific problem and solution, the idea of making the problem solving and the knowledge involved explicit, is a general one and should always be done be that based on Heuristic Classification [7], KADS [26], Generic Tasks [6] or another method to model knowledge and problem solving methods [13]. We have used the heuristic classification because it is simple, yet powerful enough.

But is this approach more than just proper system design? Well, not really. However, it does matter at what level we design, experiment with, modify and understand a system, and take advantage of various kinds of knowledge. This is crucial if we work on complex systems supporting learners where it surely is not good enough if "it works." We also must understand *why* it works.

Acknowledgments

This material is based upon work supported in part by the NSF under Grants #0437660 (IERI) and #0434624 (NSDL). Opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] R. Agrawal, T. Imielinski, and A. A. Swami. Mining associations between sets of items in large databases. In *Proceedings of the ACM SIGMOD Int'l Conference on Management of Data*, pages 207–216,, 1993.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94)*, pages 487–499, 1994.

[3] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In *Proceedings ACM SIGMOD International Conference on Management of Data*, pages 255–264, 1997.

[4] P. Brusilovsky. *Methods and Techniques of Adaptive Hypermedia*, volume 6, pages 87–129. Kluwer Academic Publishers, 1996.

[5] P. Brusilovsky, E. Schwarz, and G. Weber. ELM-ART: An intelligent tutoring system on world wide web. In C. Frasson, G. Gauthier, and A. Lesgold, editors, *Intelligent Tutoring Systems (Lecture Notes in Computer Science)*, volume 1086, pages 261–269. Springer Verlag, Berlin, 1996.

[6] B. Chandrasekaran, M. C. Tanner, and J. R. Josephson. Explaining control strategies in problem solving. *Expert Systems in Government*, pages 9–24, 1989.

[7] W. J. Clancey. Heuristic classification. *Artif. Intell.*, 27(3):289–350, 1985.

[8] K. V. Deemter, E. Krahmer, and M. Theune. Real versus template-based natural language generation: A false opposition? *Comput. Linguist.*, 31(1):15–24, 2005.

[9] C. García, Enrique Romero, S. Ventura, and T. Calders. Drawbacks and solutions of applying association rule mining in learning management systems. In *Proceedings of the International Workshop on Applying Data Mining in e-Learning (ADML'07)*, pages 14–22, 2007.

[10] P. Hansen and N. Mladenovic. J-Means: A new local search heuristic for minimum sum-of-squares clustering. *Pattern Recognition*, 34(2):405–413, 2001.

[11] R. Hübscher. What's in a prerequisite. In *Proceddings of the International Conference on Advanced Learning Technology (ICALT 2001)*, pages 165–168, Madison, 2001.

[12] A. Merceron and K. Yacef. Revisiting interestingness of strong symmetric association rules in educational data. In *Proceedings of the International Workshop on Applying Data Mining in e-Learning (ADML'07)*, pages 3–12, 2007.

[13] E. Motta. *Reusable Components for Knowledge Modelling: Case Studies in Parametric Design Problem Solving*. IOS Press, Amsterdam, The Netherlands, The Netherlands, 1999.

[14] E. R. Omiecinski. Alternative interest measures for mining associations in databases. *IEEE Transactions on Knowledge and Data Engineering*, 15(1):57–69, 2003.

[15] A. S. Palincsar. Keeping the metaphor of scaffolding fresh–a response to c. addison stone's "the metaphor of scaffolding: Its utility for the field of learning disabilities". *Journal of Learning Disabilities*, 31(4):370–73, 1998.

[16] S. Puntambekar. Analayzing navigation data to design adaptive navigation support in hypertext. In U. Hoppe, F. Verdejo, and J. Kay, editors, *Artificial Intelligence in Education: Shaping the future of learning through intelligent technologies*, pages 209–216. IOS Press, 2003.

[17] S. Puntambekar and R. Hübscher. Tools for scaffolding students in a complex learning environment: What have we gained and what have we missed? *Educational Psychologist*, 40(1):1–12, 2005.

[18] S. Puntambekar and A. Stylianou. Designing metacognitive support for learning from hypertext: What factors come into play? In U. Hoppe, F. Verdejo, and J. Kay, editors, *Artificial Intelligence in Education: Shaping the future of learning through intelligent technologies, workshop proceedings*. IOS Press, Amsterdam, 2003.

[19] S. Puntambekar, A. Stylianou, and R. Hübscher. Improving navigation and learning in hypertext environments with navigable concept maps. *Human-Computer Interaction*, 18(4):395–428, 2003.

[20] E. Reiter and R. Dale. *Building Natural Language Generation Systems.* Cambridge University Press, 2000.

[21] R. W. Schvaneveldt, editor. *Pathfinder Associative Networks: Studies in Knowledge Organization*. Ablex, Norwood, NJ, 1990.

[22] R. J. Spiro, P. J. Feltovich, M. J. Jacobson, and R. L. Coulson. Cognitive flexibility, constructivism, and hypertext: Random access instruction for advanced knowledge acquisition in ill-structured domains. *Educational Technology*, May:24–33, 1991.

[23] C. A. Stone. The metaphor of scaffolding: Its utility for the field of learning disabilities. *Journal of Learning Disabilities*, 31(4):344–364, 1998.

[24] T. A. Van Dijk and W. Kintsch. *Strategies of discourse comprehension.* New York: Academic, 1983.

[25] L. S. Vygotsky. *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge, MA, 1978.

[26] B. J. Wielinga, A. T. Schreiber, and J. A. Breuker. Kads: a modelling approach to knowledge engineering. *Knowl. Acquis.*, 4(1):5–53, 1992.

[27] D. Wood, J. S. Bruner, and G. Ross. The role of tutoring in problem solving. *Journal of Child Psychology & Psychiatry & Allied Disciplines*, 17(2):89–100, 1976.