

Rough Set Approach to User Modeling

Binghui Helen Wu
FS Consulting
Wichita, Kansas
Swhelen21@hotmail.com

Abstract—The intelligence of an interactive system is exhibited through the evolution of its intelligent behaviors, namely, its effectiveness in performing tasks and its adaptiveness to both variations of task conditions and different users. Having a good user modeling component in any intelligent interactive system becomes crucial to the performance of that system. We have developed a computational approach to building intelligent tutoring systems (ITS) in conjunction with our framework for ITS [1]. This computational approach employs two exploratory AI techniques—a hybrid knowledge representation model as an implementation mechanism of ITS and a rough set [2] approximation approach to user modeling. In this article, we focus on discussing the rough set approach to user modeling of the system in detail.

Keywords—rough sets; user modeling; computational approach to knowledge representations; intelligent tutoring systems

I. INTRODUCTION

Research in building intelligent computer assisted learning environments (more often known as intelligent tutoring systems, ITS) has been ongoing for more than 20-25 years. Existing ITS research, including well-known tutoring systems such as Clancey's GUIDON and its revision [3], Anderson's ACT Tutors [4], Johnson's PROUST [5], and many others, have shown that the key feature of tutoring systems is their ability to diagnose students' misconceptions in learning during a tutoring process and, based upon those diagnoses, to provide students with remedial learning. However, many psychological questions underlying learning, teaching and understanding have not received convincing answers. In addition, enormous difficulties in creating a precise machine representation of the students' learning stages and in identifying students' possible misconceptions lead to a diversification of ITS research. Many urged the importance of evaluation of such systems, for example, the special issue on evaluation in *IJ. of Artificial intelligence in Education* (1993) 4(2/3).

In spite of difficulties in existing ITS research, we believe that tutoring by computers is an attractive alternative to human tutors in providing assistance, both to instructors who are responsible for students' learning and to individual students who are supposed to learn the subject to be taught, especially when human tutors are not available or not willing to spend more time. In fact, one study has shown [6], tutoring by a human instructor—a one-on-one help session to improve an individual student's learning—is more effective than that in a classroom instruction setting.

Aiming at building an effective tutoring system, which is as adaptable to different students and instructors as possible, we have proposed a framework (ATA, i.e., Artificial Teaching Assistant) for ITS [1] based on the constructivist theory of

knowledge, learning, and understanding. Constructivism is a model of how learning takes place. It has drawn considerable attention among science education researchers and ITS researchers who aim at building computer based learning environment [7] [8] [9]. In the constructivist view, learning is an active process of constructing a conceptual framework. Although sharing the basic view of learning, different interpretations of this view result in different approaches to science education. Our understanding of constructivist theory is based on Cobern's interpretation [10], in which our ATA framework for ITS reflects.

This ATA framework consists of three models—presentation, question, and test models—attempting to capture the two most important aspects of a tutoring process. One of the two aspects is how the presented knowledge is organized, so that it can be presented automatically and effectively. The other is how questions and tests are to be given to ensure that students have developed a meaningful interpretation of the presented knowledge and can apply it to solve related problems. To exemplify our ATA framework for ITS, we have been implementing a prototyped tutoring system with number systems, a topic in Foundations of Computer Science [11], as its domain. A systematic approach has been developed for the implementation. In this paper, we first give an overview of our computational approach to building ITS, in which a hybrid knowledge representation model (SG, i.e., Graph-Sets) provides both an automated presentation process of domain knowledge and automated interpretation of student learning and/or understanding of the presented knowledge. Then, we present a novel approach to user modeling—rough set approximation—that we have used in the user modeling component of the implementation of our tutoring system.

II. OVERVIEW OF OUR COMPUTATIONAL APPROACH TO ITS

Our hybrid representation model (GS model), which mixes two levels of knowledge representation schemes, subject network—a fully connected, directed, and acyclic graph with a semantic interpretation and "knowledge base"—a two tuple, a universal discourse and a class of relations imposed on the universal set. There are two types of knowledge possessed by the paradigm of our tutoring system: the knowledge of the subject materials to be represented and the knowledge of how to make an effective presentation dynamically based on various student responses during tutoring so that the quality of tutoring can be ensured. Based on these, we use a variation of semantic networks (named subject network), as the coarse grain of representation of domain knowledge; and we use "knowledge base" (coined by Pawlak, see the chapter in [2]), as the fine grain representation of the domain knowledge.

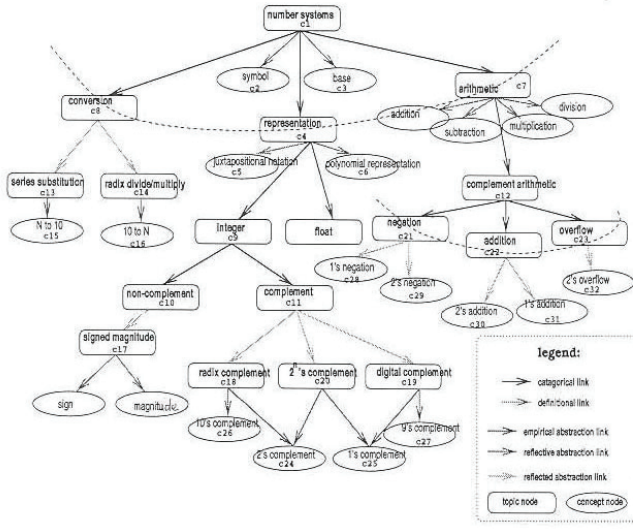


Figure 1: An Example to illustrate the ATA representation model of our framework for ITS (see Appendix for a clear version)

A. Knowledge Representation Scheme I: Subject Network

In semantic networks, knowledge is represented in terms of nodes, which can represent objects, and are connected by links that represent the interrelations among nodes. This graphical form of knowledge representation is most appealing and has long been used in AI systems including ITS [12] [13]. In the context of our tutoring system where abstract knowledge has been chosen as the domain of tutoring materials, naturally, we have adopted this knowledge representation scheme. However, there are two important features that have distinguished our way of using this scheme.

First, as shown in Figure 1, there are two different kinds of nodes: topic and concept, in which knowledge of the subject material has been stored. There are also two kinds of links to capture two kinds of relations among nodes: inclusion and dependency. An inclusion link describes the relation between two nodes in which the destination node always elaborates its source node (i.e., is a subtopic of the topic represented in the source node). On the other hand, a dependency link describes a dependency relation between two nodes with two possibilities. If a topic node depends on a concept node, the concept is used to define the topic node; and, if a dependency link exists between two topic nodes, the source node would be the prerequisite topic of the destination node. For example (see Figure 1), the topic, *number systems*, can be defined using concepts, *symbol* and *base* (dependency) and can be elaborated with subtopics, *representation*, *arithmetic*, and *conversion* of numbers with different bases (inclusion). Further, in our number systems example, in order to introduce the subtopic *conversion* (between numbers with different bases), we need to have introduced how a number is represented (*representation*).

Therefore, the relation between number representation and number conversion is a dependency relation. Also, note that both topic and concept nodes are used to store knowledge in a subject network; they are both literally concepts. The difference between the two, besides their names, is that the concept in a concept node in a subject network is atomic in the sense that it doesn't need to be further elaborated. Therefore, concept nodes are minimal elements of subject graphs.

Second, in addition to the nodes and links described above, special kinds of nodes depicted in the dashed notations and named supplementary nodes, are utilized in the subject graph to represent supplementary material as options for students who desire to know more and learn further about the subject. A supplementary node may be either a topic or concept node. Continuing with our example, since the subtopic *conversion* mainly provides a better understanding of numbers, it may not be necessary to teach this subtopic if we present the topic, *number systems*, in connection with digital computers. Thus, we define the subtopic as a supplementary material to students' learning of number systems.

Overall, the knowledge representation scheme described above provides not only a way of establishing a network of subject material, but also an automatic presentation procedure: that is, an automatic scheme for traversing nodes by following links which may be optionally set by a user (an instructor). Now the problem left to be investigated is how can we ensure that a student is learning while following through a presentation? We believe that the behaviorist instructional teaching [14] can be effective. Specifically, we use Gagne's nine-significant-event model—recursively utilizing the nine events while instructing: gaining learner's attention, informing learner of objectives, stimulating the recall of prerequisites, presenting stimuli, providing guidance, eliciting performance, providing feedback, assessing performance, and enhancing retention and transfer. Overall, we ask questions while presenting materials and give exercise problems as homework so that we can assess students' performance in their learning. Then, the question is how can we implement this kind of tutoring so that our system can do it effectively?

B. Knowledge Representation Scheme II: Knowledge Base

Pawlak has proposed a formal definition of knowledge in terms of Knowledge Base and its basic property [2]. Directly motivated by his understanding and formalization of "knowledge" and "concepts," we employ his set notations, known as Knowledge Base, as a fine grain representation of the presented knowledge in tutoring systems.

In our representation, concepts in a subject network of a tutoring system used to evaluate students' learning are encoded in questions and problems, which are accessed through interactions between students and the system during tutoring. That is, those of concepts are used for questioning and testing students, and evaluating students' answers. For instance, in our number systems example, there are about 35 concepts, including 11 supplementary concepts used for testing and evaluating students' learning of the subject material; and there are about 50 questions and problems that have been stored in the associated database of the tutoring system, which may be

assigned to students as exercises, where each problem embeds several concepts among the fifty. In general, let

$$C = \{c_1, c_2, \dots, c_n\} \quad (1)$$

be the set of n concepts that are presented in a subject network. Further, let there be m problems in the problem set

$$P = \{p_1, p_2, \dots, p_m\} \quad (2)$$

that have included these n concepts in the subject network. One way of representing such knowledge is to use n by m matrix as the following.

	p_1	p_2	p_3	...	p_m
c_1	+	+			
c_2			+		
c_3					+
...					
c_n	+				+

In this matrix, a cell is checked means that the concept identified by the row of that cell is embedded in the problem identified by the column of that cell. Further, a student's answer to the problem can be checked against these concepts. We then keep track of students' answers to problems, which should be a reflection of their understanding of those concepts.

Although we have found a method to represent subject material (the use of GS-model) and student learning stages (eliciting performance of student's learning of each concept) mechanically and to test and evaluate students' learning through their answering questions and solving problems, there are many possible kinds of uncertainties inherent in this paradigm of tutoring.

III. USER MODELING TO COPE WITH UNCERTAINTIES

We have identified two kinds of uncertainties, namely, vagueness and incomplete information. Vagueness may exist in the methods of evaluating students' answers and in the answers that students give to problems. First, as to the methods of evaluating students' answers, it is likely that relations between concepts may not be simple, but instead, hierarchical. For example, in the subject domain of our tutoring system, complement is an important number representation in digital computers, which provides uniform computation for arithmetic operations of both positive and negative numbers. There are two variations of this representation, radix and digit complements. Complement, radix complement, and digit complement are considered to be different concepts in our representation model. Obviously, they are closely related. That is, one's understanding of radix complement may imply his understanding of complement and partial understanding of digit complement. Generally depending on how an instructor would view these related concepts, the subject material in a tutoring system could be rearranged with significant differences, so can problems to test students' mastery level to those concepts.

Second, as to the students' answer, again according to the behaviorist model of understanding students' learning, if a students' answer to a problem is incorrect, generally, we infer that the student has not yet mastered the concept(s) that the problem tries to test. However, the information on the cause of students' wrong answers is somewhat incomplete (perhaps, it would be impossible to obtain it fully), and may result from students' psychological factors, their mental states during a tutoring session. Simply presenting the tutoring material again to students may not be an effective remedial learning. Instead, testing students with more problems and approximating their answers could be a better solution. This means that more flexible ways of evaluating students' answers may be desirable or even necessary.

Motivated by seeking how to build an accurate user model, we have explored the rough set formalism and adopted it in the user-modeling component of our tutoring system to compensate for these uncertainties.

A. The Rough Set Approximation

Pawlak has succinctly described [15]:

"In the rough set theory each vague concept is replaced by a pair of precise concepts called its lower and upper approximations; the lower approximation of a concept consists of all objects which surely belong to the concept, whereas the upper approximation of the concept consists of all objects which possibly belong to the concept." (p. 73 in [2])

Corresponding to his description of vague concept, we have here the vagueness of evaluating students' learning and the vagueness of understanding of the relations between concepts depending upon instructors' preference.

Formally, let U be a set of universe of discourse (called universe); let R be a set of equivalence relations imposed on the universe U .

The Knowledge Base K is defined as $K = (U, R)$; and the concept X in K is defined as a subset of U , i.e., $X \subseteq U$.

Further, the upper approximation of the concept X in K under a given equivalence relation R is defined as

$$\overline{X}_R = \{x : [x]_R \cap X \neq \emptyset\} \quad (3)$$

and the lower approximation of the concept X is defined as

$$\underline{X}_R = \{x : [x]_R \subseteq X\} \quad (4)$$

Here $[x]_R$ represents the equivalence class of x under the given relation R . If $\overline{X}_R = \underline{X}_R$, then the concept X is considered to be precise; otherwise, we say the concept is vague. For a vague concept, these two approximations under the given relation in the Knowledge Base K can be calculated and obtained.

For example, as we have previously written, let (1) be the set of n concepts that are presented in a subject network. An interrelationship among concepts can be expressed as an equivalence relation (a partition). Then, there are many

possible partitions to form equivalence classes on the set C . Further, because there are m problems in the problem set (2) that have included these n concepts in the subject network, there should be various combinations for an instructor to assign a student problems according to the instructor's preference (expressed in his way of partitioning concepts in the subject network) and using this partition, problems are assigned. Further, an evaluation of the students' answers involving each concept can be obtained. Among the possible answers that a student may give, an evaluation based on the rough set approximation may be performed.

In our tutoring system, specifically, in one of the subject networks of the tutoring system, we can presumably represent all concepts that we would like students to learn with the possibility of varying granularity level of each node; and we also can select sufficient problems to embed these concepts, and for students to solve. Formally, let P_i be a subset of the problem set P , that is,

$$P = \cup P_i, i \in 1, 2, \dots, m \quad (5)$$

in the subject network. Let C_i be a subset of the concept set C , that is,

$$C = \cup C_i, i \in 1, 2, \dots, n \quad (6)$$

in the subject network. We can impose relations between P and C functionally as

$$f_{P_i} : P_i \rightarrow C \text{ where } i \in 1, 2, \dots, l \quad (7)$$

and each of these l functions is subjective; and

$$f_{C_i} : C_i \rightarrow P \text{ where } i \in 1, 2, \dots, k; \quad (8)$$

and it (f_{C_i}) indicates that, given a group of concepts, a corresponding set of problems can be selected to test students' learning of this group of concepts. PC (Note that $P_i \in 2^P$ and $C_i \in 2^C$.) Providing the mechanism based on the rough set theory, our system allows instructors to group concepts (C_i) and to select problems (P_i) differently so that the evaluation of students' answers to problems can vary according to instructors' preferences. Our objective here is to find good selection algorithms such that the vagueness in ensuring and evaluating students' learning in the tutoring processes will be minimized.

B. Comparison with Other Approximation Approaches

Katz et al.[16] have proposed using fuzzy probabilistic variables to measure performance conditions when students are learning with their tutoring system. More recently, Kavcic [19] has conjectured with fuzzy and probabilistic user modeling. In fact, probability theory [17] and fuzzy set theory [18] are popularly used in approximate reasoning. However, the advantage of the rough set approximation over these two in our case is that the numerical value of imprecision is

calculated according to a pre-defined formula which is based on the fundamental concepts used to express imprecision of knowledge. For example, Pawlak has proposed one numerical calculation of the rough set imprecision, namely,

$$a_R = \frac{\text{card } \underline{R}}{\text{card } \overline{R}} \quad (9)$$

where card denotes the cardinality of sets ($\overline{R}$ or $\underline{R}$). To illustrate this point, we use a simple example as follows. Let there be 10 concepts that have been chosen to be presented to students. One instructor may partition these 10 concepts as

$$[c_1]_R = \{c_1, c_2, c_3\}, \quad (10)$$

$$[c_4]_R = \{c_4, c_5\}, \quad (11)$$

$$[c_6]_R = \{c_6\}, \quad (12)$$

$$[c_7]_R = \{c_7, c_8\}, \quad (13)$$

$$[c_9]_R = \{c_9, c_{10}\}, \quad (14)$$

That means that concepts in the same equivalent class are substitutable. Suppose one student has mastered the following concepts:

$$\{c_1, c_2, c_4, c_6, c_7, c_9, c_{10}\}; \quad (15)$$

Then the upper approximation of his performance under the given relation is

$$\{c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8, c_9, c_{10}\}; \quad (16)$$

while the lower approximation of his performance is

$$\{c_6, c_9, c_{10}\} \quad (17)$$

Consequently, the accuracy using Pawlak's calculation would be $2/5$. On the other hand, the fuzzy probability variables in [16] are obtained by expert estimations without a predefined membership function. As a result, the inconsistency of human performance would bias the approximation processes inevitably.

IV. CONCLUSION AND FURTHER INDICATION

We have developed the rough set approach to user modeling under our ATA framework for ITS and our hybrid knowledge representation model (GS model) and implemented it in the user modeling component of our tutoring system. Theoretically, our development approaches are based on the constructivist understanding of scientific knowledge [9] [10] and the constructivist view of learning mathematical knowledge [8]. After we have formulated our hybrid representation model (GS model), the rough set approximation

fits well in the implementation of our tutoring system. As we have briefly described, under the ATA framework, GS model provides an automated presentation mechanism (both initial and remedial presentations). Thus, the rough set approach can be used on top of the model to make the tutoring process more adaptable.

Our intention has been using exploratory AI approaches to build a tutoring system with observable intelligent behavior in its performance. Other important factors confronted in our implementation have been interface design issues. These issues may contribute to user modeling, and could be crucial to the performance of our system as well. As we have tried to show in this article, our automation mechanism is based on the high consensus approach—i.e., providing questions with multiple choices, and exercises that have solutions with well-defined steps— in evaluating students' learning. Wasilewska has recently formulated Knowledge Systems [20] based on Pawlak's Knowledge base in the context of data mining. It can also be applied to the user modeling component of our ITS. We strongly believe that with these exploratory AI approaches—

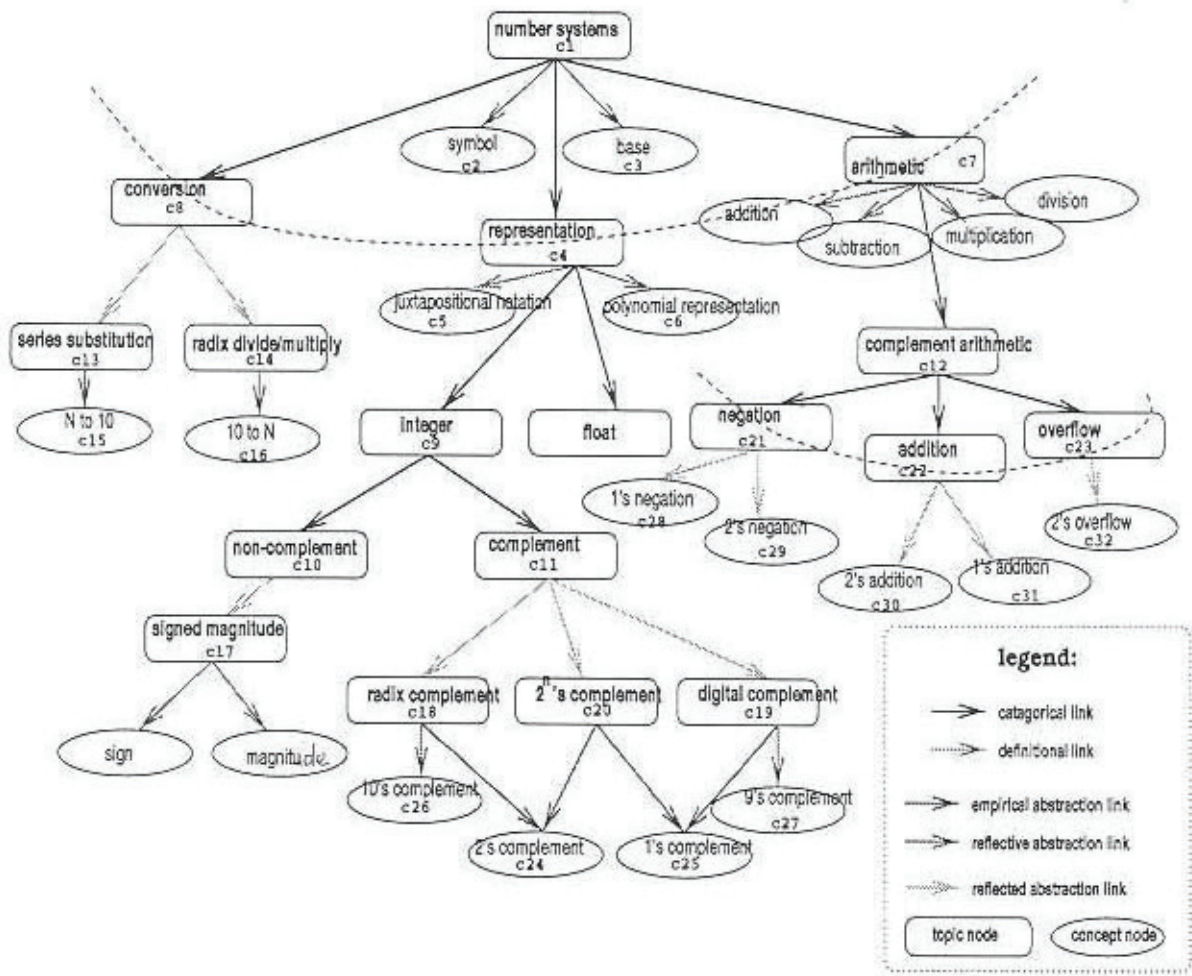
i.e., the hybrid representation model and the rough set approximation—our tutoring system has had a lot to offer to ITS developers and will have a significant potential in the development of ITS.

ACKNOWLEDGMENT

Thanks to the Electrical Engineering and Computer Science Department of Lehigh University for the financial support of this work. Thanks also to T. Boulton and G. Blank, associate professors of EECS Department of Lehigh University for their support, especially for providing criticism on this work.

APPENDIX

Figure 1 in large size



REFERENCES

- [1] Binghui Helen Wu. A framework for intelligent tutoring systems. In The 7th World Conference on Artificial Intelligent in Education (AI-ED 95), page 599. AACE and GMU, Aug. 1995.
- [2] Zdzislaw Pawlak. Rough Sets: Theoretical Aspects of Reasoning about Data. Kluwer Academic, The Netherlands, 1991.
- [3] William J. Clancey. Guidon-manage revisited: A socio-technical systems approach. *Jl. of Artificial Intelligence in Education*, vol.4, no. 1 p5-34, 1993.
- [4] John R. Anderson. Cognitive tutors: Lessons learned. In The 4th International Conference on User Modeling, page 71. MITRE Corp., Aug. 1994.
- [5] W. Lewis Johnson. Understanding and debugging novice programs. *Artificial Intelligence*, vol. 42, p51-97, 1990.
- [6] John R. Anderson and Brian J. Reiser. The lisp tutor. *BYTE*, vol. 4, p159-175, 1985.
- [7] Warren Sack, Elliot Soloway, and Peri Weingrad. Re: Writing Cartesian student models. *Jl. of Artificial Intelligence in Education*, vol. 3, no. 4, p381-399, 1992..
- [8] Paul Cobb, Erna Yackel, and Terry Wood. A constructivist alternative to the representational view of mind in mathematics education. *Journal for Research in Mathematics Education*, vol. 4, no.1,p2-33, 1992.
- [9] John P. Smith, Andrea A. diSessa, and Jeremy Roschelle. Misconceptions reconceived: A constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences*, vol. 3, no. 2, p115-164, 1993/94.
- [10] J. William W. Cobern. Constructivism. *Journal of Educational and Psychological Consultation*, vol. 4, no.1, p105-112, 1993.
- [11] Alfred V. Aho; Jeffrey D. Ullman. *Foundations of Computer Science*. Computer Science Press, an Imprint of W. H. Freeman and Company, New York, 1992.
- [12] Renee Elio and Peternela B. Scharf. Modeling novice-to-expert shifts in problem solving strategy and knowledge organization. *Cognitive Science*, vol. 14, no.4,p 579-639, 1990..
- [13] Denise Whitson Gurer. A Bi-level Physics Student Diagnostic Utilizing Cognitive Models for an Intelligent Tutoring System. PhD thesis, Lehigh University, Aug. 1993.
- [14] Robert M. Gagne, Leslie J. Briggs, and Walter W. Wager. *Principles of Instructional Design*. Harcourt Brace Jovanovich College Publishers, Orlando, FL, 1992.
- [15] Zdzislaw Pawlak. Rough sets present state and further prospects. In The Third International Workshop on Rough Set and Soft Computing (RSSC '94), pages 72-76. AAAI, EPRI, IAMCS, ISF, IEEE, and SJSU, Nov. 1994
- [16] Sandra Katz, Alan Lesgold, Fary Eggan, and Maria Gordin. Modelling the student in sherlock ii. *Jl. of Artificial Intelligence in Education*, vol. 3, no.4,p 495-518, 1992..
- [17] P. Cheeseman. In defense of probability. In *Proceedings IJCAI-85*, 1985.
- [18] L.A. Zadeh. A theory of approximate reasoning. In *Machine Intelligence*. Halstead Press, New York, 1979
- [19] Alenka Kavcic. Fuzzy User Modeling for Adaptation in Educational Hypermedia. *IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews*, vol. 34, no. 4, November 2004
- [20] Anita Wasiewska. Descriptive Granularity. 2010 IEEE International Conference on Granular Computing, San Jose, California, 14-16 August 2011