Combination of Bayesian Network and Overlay Model in User Modeling

Loc Nguyen[1] and Phung Do[2]

[1] University of Natural Science / Faculty of Information Technology,
Ho Chi Minh City, VietNam
[2] University of Information Technology / Faculty of Information System,
Ho Chi Minh City, VietNam
ng_phloc@yahoo.com

Abstract. The core of adaptive system is user model containing personal information such as knowledge, learning styles, goals which is requisite for learning personalized process. There are many modeling approaches, for example: stereotype, overlay, plan recognition… but they do not bring out the solid method for reasoning from user model. This paper introduces the statistical method that combines Bayesian network and overlay modeling so that it is able to infer user's knowledge from evidence collected during user's learning process.

Keywords: Bayesian network, overlay model, user model.

1 Introduction

User modeling is the new trend of enhancing the adaptability of e-learning system. User models are classified into: stereotype model, overlay model, differential model, perturbation model, plan model.

- Stereotype [27] is a set of user's frequent characteristics. In general, stereotype represents a category or group of learners.
- In overlay modeling, learner model is the subset of domain model. The domain is decomposed into a set of elements and the overlay model is simply a set of masteries over those elements.
- Differential model is basically an overlay on expected knowledge, which in turn is an overlay on expert's domain knowledge.
- Perturbation model represents learners as the subset of expert's knowledge plus their mal-knowledge.

Modeling user must follow three below steps:

- Initialization is the process that gathers information and data about user and constructs user model from this information.
- Updating intends to keep user model up-to-date.
- Reasoning new information about user out from available data in user model.

Reasoning is complex but essential and interesting, especially, there is need to deal with uncertain or imprecise information in user modeling. For example, answering the question: "The student failed the exam, so most probably he doesn't master the knowledge" is involved in processing uncertain information. The approaches which solve this problem primarily base on theory of artificial intelligence (AI) or statistics. Both AI and statistics have particular advantages and drawbacks but statistical method is appropriate to evaluate learner's performance by collecting evidence. Bayesian network which is the marriage between Bayesian inference and graph theory has a solid mathematical fundamental. Additionally, overlay model can represent very clearly user's knowledge.

In this paper, we propose the combination of overlay model and Bayesian network so that it is able to take full advantages of strong points of both of them.

Section 2: survey of Bayesian inference and Bayesian network, the core of our method. Section 3: Applying Bayesian network to overlay model. Section 4: Evaluation of this method and conclusion.

2   Bayesian Network

2.1 Bayesian Rule

Bayesian inference, a form of statistical method, is responsible for collecting evidence to change the current belief in given hypothesis. The more the observed evidence, the higher degree of belief in hypothesis is. First, this belief is assigned an initial probability. Note, in classical statistical theory, the random variable's probability is objective (physical) through trials. But, in Bayesian method, the probability of hypothesis is "personal" because its initial value is set subjectively by expert. When evidence is gathered enough, the hypothesis is considered trustworthy.

Bayesian inference is based on Bayesian rule with some special aspects:

$$P(H \mid E) = \frac{P(E \mid H) * P(H)}{P(E)} \tag{1}$$

H is probability variable denoting a hypothesis existing before evidence
E is also probability variable notating an observed evidence
P(H) is *prior probability* of hypothesis. It is also hypothesis' initial value
P(H | E), conditional probability of H with given E, is called *posterior probability*. It tell us the changed belief in hypothesis when occurring evidence
P(E | H) is conditional probability of occurring evidence E when hypothesis is true. In fact, likelihood ratio is P(E | H) / P(E) but P(E) is constant value. So we can consider P(E | H) as *likelihood function* of H with fixed E.
P(E) is probability of occurring evidence E together all mutually exclusive cases of hypothesis. If   H and E are discrete, $P(E) = \sum_{H} P(E \mid H) * P(H)$, otherwise $f(e) = \int f(e \mid h) f(h) dh$ with h and e being continuous, f denoting probability density function. Because of being sum of products of prior probability and likelihood function, P(E) is called marginal probability.

Note: H, E must be random variables according to statistical theory.

2.2   Bayesian Network

Bayesian network is combination of graph theory and Bayesian inference. It having a set of nodes and a set of directed arcs is the directed acyclic graph (DAG). Each node represents a random variable which can be a evidence or hypothesis in Bayesian inference. Each arc reveals the cause-effect relationship among two nodes. If there is the arc from node A to B, we call "A causes B" or "A is the parent of B", in other words, A depends conditionally on B. Otherwise there is no arc between A and B, it asserts the conditional independence. Note, in Bayesian network context, terms: node and variable are the same.

A node has a local conditional probability distribution (CPD). If variables are discrete, CPD is simplified as table (CPT). When one node is conditionally dependent on another, there is a corresponding probability (in CPT or CPD) measuring the influence of causal node on it. In case node has no parent, its CPT degenerate into prior probabilities.

For example, in figure 1, event "cloudy" is cause of event "rain" or "sprinkler", which in turn is cause of "grass is wet". So we have three causal relationships of: 1-cloudy to rain, 2- rain to wet grass, 3- sprinkler to wet grass. This model is expressed below by Bayesian network with four nodes and three arcs corresponding to four events and three relationships. Every node has two possible values True (1) and False (0) together its CPT.
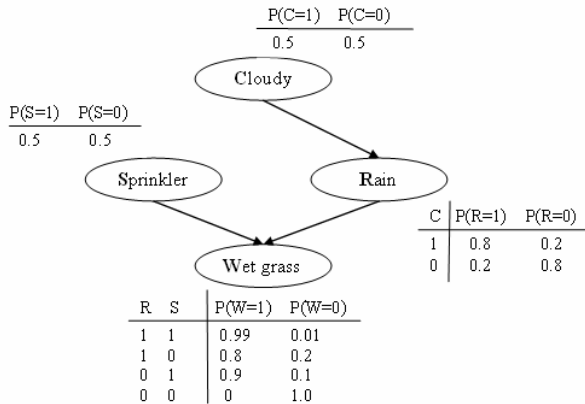

Fig. 1. Bayesian network (a classic example about "wet grass")

Suppose we use two letters x_i and $pa(x_i)$ to name a node and a set of its parent, correspondingly. X is vector which was constituted of all x_i, $X = (x_1, x_2,..., x_n)$. The global joint probability distribution $p(X)$ being product of all local CPDs or CPTs is formulated as:

$$p(x1, x2,..., xn) = \prod_{i=1}^{n} p(xi \mid pa(xi))$$

(2)

Suppose Ω_i is the subset of $pa(x_i)$ such that x_i must depend conditionally and directly on every variable in Ω_i. In other words, there is always an arc from each variable in Ω_i to x_i and no intermediate node between them.

Thus, formula 2 becomes:

$$p(x1, x2, ..., xn) = \prod_{i=1}^{n} p(xi \mid \Omega i) \qquad (3)$$

In figure 1, according to formula 2:

$$p(C, R, S, W) = p(C)p(R \mid C)p(S \mid C)p(W \mid C, R, S)$$

Applying formula 3, $p(S \mid C) = p(S)$ due to the conditional independence assertion about variables S and C. Furthermore, because S is intermediate node between C and W, we should remove C from p(W | C, R, S), hence, $p(W \mid C, R, S) = p(W \mid R, S)$. In short, the expansion of formula 3 is shown below:

$$p(C,R,S,W) = p(C)p(S)p(R \mid C)p(W \mid R,S) \qquad (4)$$

2.3   Inference in Bayesian Network

Using Bayesian reference, we need to compute the posterior probability of each hypothesis node in network. In general, the computation based on Bayesian rule is known as the inference in Bayesian network.

Reviewing figure 1, suppose W becomes evidence variable which is observed the fact that the grass is wet, so, W has value 1. There is request for answering the question: how to determine which cause (sprinkler or rain) is more possible for wet grass. Hence, we will calculate two posterior probabilities of S (=1) and R (=1) in condition W (=1). These probabilities are also called *explanations* for W.

$$p(R = 1 \mid W = 1) = \frac{\sum_{C,S} p(C, R = 1, S, W = 1)}{\sum_{C,R,S} p(C, R, S, W = 1)} \qquad (5)$$

$$p(S = 1 \mid W = 1) = \frac{\sum_{C,R} p(C, R, S = 1, W = 1)}{\sum_{C,R,S} p(C, R, S, W = 1)} \qquad (6)$$

In fact, formulas 5 and 6 are expansion of formula 1. Applying (4) to (5) & (6):

p(R=1|W=1)=0.4475/0.7695=0.581<p(S=1|W=1)=0.4725/0.7695=0.614

It is concluded that sprinkler is the most likely cause of wet grass.

3   Applying Bayesian Network to Overlay Model

The basic idea of overlay modeling is that the user model is the subset of domain model. Straightforward, the domain is decomposed into a set of knowledge elements and the overlay model (namely, user model) is simply a set of masteries over those elements. Suppose that the mastery of each element varies from 0 (*not mastered*) to

1 (*mastered*), to wit weighted overlay. The relationship of element A to element B is often prerequisite relationship, so, we can deduce that user must comprehend A before learning B. Then the expert model is the overlay with 1 for each element and the learner model is the overlay with at most 1 for each element.

Although overlay model is the simple but powerful method to represent user model, it does not provide the way to infer user's knowledge from evidence collected in user's learning process. Overlay modeling should associate with other statistical approach in solving this problem and Bayesian network is the best choice. So, we combined Bayesian network and overlay model by following steps:

1. The structure of overlay model is considered as Bayesian network. Thus, knowledge elements in domain become variables (or nodes) in Bayesian network. Instead of using the weight of each element as above, we assign the probability to each variable for estimating the mastery of knowledge. All variables are binary (0 – not mastered and 1 – mastered). Note, knowledge item, knowledge element and concept are synonymical terms.
2. The prerequisite relationships between knowledge elements are known as the conditional dependence assertions in Bayesian network. Accordingly, each node has a CPT.
3. All knowledge elements are defined as *hidden variables* (hypothesis). Other learning objects or events (such as: tests, exams, exercises, user's feedback, user's activities) which are used to assess or evaluate user's performance in learning process are consider as *evidence variables*. We must add them to Bayesian network along with determining the conditional dependence relationship between them and remaining hidden variable, namely, specifying their CPTs. Inferring user's knowledge is to compute posterior probability of hidden variables (according to formula 5, 6) when evidence variables change their values. This process can be known as knowledge diagnosis.
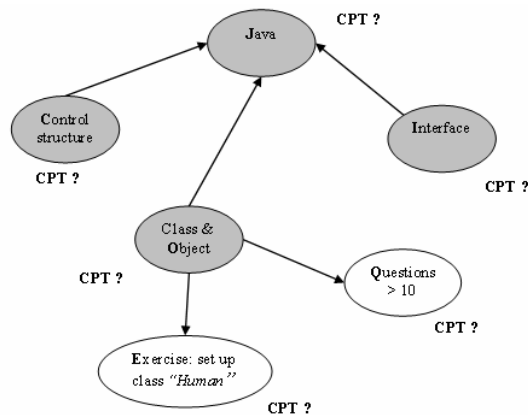
Fig. 2. Combination of Bayesian network and overlay model *(evidence nodes are unshaded, otherwise, hidden nodes are shaded)*

After completing the three steps described above, creating the learning network leaves us with two essential conditions:

- Specifying the structure of model including nodes and arcs. This task is development of qualitative model done by experts such as teachers, lecturers, supervisors… or by learning algorithms.
- Specifying the important parameters which are CPTs of all variables. This task called development of quantitative model is described in section 3.1.

3.1 Specifying CPTs of Variables

Suppose Java course is constituted of four concepts considered as hidden variables whose links are prerequisite relationships. Additionally, there are two evidence variables: "Questions > 10" and "Exercise: set up…". That learner asks more than 10 questions is to tell how much her/his amount of knowledge. Like that, evidence "Exercise: set up…" proves whether or not he/she understands concept "Class & Object". The number (in range 0…1) that measures the relative importance of each prerequisite or evidence is defined by expert or teacher. In other words, this is the *weight* of arc from parent node to child node. All weights concerning the child variable will build up its CPT. Sum of weights of all arcs to/from each child/parent node in case of hidden/evidence variable should be 1. It means that each weight is normalized.

Your attention please, the relationship between hidden variable (H) and evidence variable (E) must be from H to A because the process that computes posterior probability of hidden variable with evidence is the knowledge diagnosis. So, evidence variable has no child and its parents must be hidden variables. In short, there are two kinds of relationships:

- Prerequisite relationships among hidden variables.
- Diagnostic relationships of hidden variables to evidence. The mastery of concepts (hidden) effects on the trust of evidence. However, if learner failed an examination, it is not sure about her/his lack of knowledge or ability because she/he can make a mistake unexpectedly.
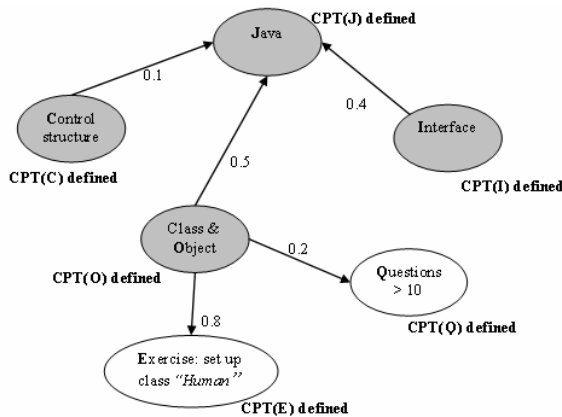
Fig. 3. Bayesian overlay model and its parameters in full *(evidence nodes are unshaded, otherwise, hidden nodes are shaded)*

In this example, node J (Java) has three parents: C (Control structure), O (Class & Object), I (Interface) which in turn are corresponding to three weights of prerequisite relationship: $w_1=0.1$, $w_2=0.5$, $w_3=0.4$. Conditional probability of J is computed as follows:

$$p(J \mid C, O, I) = w_1*h_1 + w_2*h_2 + w_3*h_3$$

where $h1 = \begin{cases} 1 \ if \ C = J \\ 0 \ otherwise \end{cases}$ $h2 = \begin{cases} 1 \ if \ O = J \\ 0 \ otherwise \end{cases}$ $h3 = \begin{cases} 1 \ if \ I = J \\ 0 \ otherwise \end{cases}$

Note: {J, C, O, I} is complete set of mutually exclusive variables (of course, each also variable is random and binary). Generalizing about formula 7, it is that:

$$p(X = 1 \mid Y1, Y2, ..., Yn) = \sum_{i=1}^{n} wi * hi \qquad (7)$$

where $hi = \begin{cases} 1 \ if \ Yi = X \\ 0 \ otherwise \end{cases}$ with given random binary variables X, Yi. Obviously, *p(not X | Y1, Y2,…, Yn) = 1 - p(X | Y1, Y2,…, Yn)*.

Applying formula 7, CPT of J, E, Q is determined below:

Table 1. CPTs of J, E, Q. Namely, T_1, T_2, T_3

T_1			
C O I	**p(J = 1)**	**P(J = 0)** 1- p(J = 1)
1 1 1	**1.0** *(0.1*1 + 0.5*1 + 0.4*1)*	**0.0**
1 1 0	**0.6** *(0.1*1 + 0.5*1 + 0.4*0)*	**0.4**
1 0 1	**0.5** *(0.1*1 + 0.5*0 + 0.4*1)*	**0.5**
1 0 0	**0.1** *(0.1*1 + 0.5*0 + 0.4*0)*	**0.9**
0 1 1	**0.9** *(0.1*0 + 0.5*1 + 0.4*1)*	**0.1**
0 1 0	**0.5** *(0.1*0 + 0.5*1 + 0.4*0)*	**0.5**
0 0 1	**0.4** *(0.1*0 + 0.5*0 + 0.4*1)*	**0.4**
0 0 0	**0.0** *(0.1*0 + 0.5*0 + 0.4*0)*	**1.0**

T_2		
E	**p(E = 1)**	**p(E=0)** 1-p(E=1)
1	**0.8** (0.8*1)	**0.2**
0	**0.0** (0.8*0)	**1.0**

T_3		
Q	**p(Q = 1)**	**p(Q=0)** 1-p(Q=1)
1	**0.2** (0.2*1)	**0.8**
0	**0.0** (0.8*0)	**1.0**

Because concepts C, O, I has no prerequisite knowledge for understanding, their CPTs are specified as prior probabilities obeying uniform distribution (assigned medium value 0.5 in most cases).

Table 2. CPTs of C, O, I. Namely, T_4, T_5, T_6

P(C=1)	P(C=0)		P(O=1)	P(O=0)		P(I=1)	P(I=0)
0.5	0.5		0.5	0.5		0.5	0.5

3.2 Inferring User's Knowledge

Suppose a leaner did well the exercise "Set up… " and asked more than 10 question. That is to   say the occurrence of two evidence, namely, E = 1 and Q = 1. It is necessary to answer the question: *How mastered is learner over the concept "Java"*? Thus, the posterior conditional of hidden variables J with fixed events E =1 and Q=1, *p(J = 1 | C, O, I , E = 1, Q = 1)*, must be computed.

According to formula 5, 6:

$$p(J = 1 | C, O, I, E = 1, Q = 1) = \frac{\sum_{C,O,I} p(J = 1, C, O, I, E = 1, Q = 1)}{\sum_{C,O,I,E,Q} p(J = 1, C, O, I, E, Q)}$$

where p(J, C, O, I, E, Q) is global joint probability distribution, p(J, C, O, I, E, Q) = p(C) * p(O) * p(I) * p(E|O) * p(Q|O) * p(J|C,O,I).

Applying all CPTs in table 1, 2, 3, 4, 5, 6, it is able to determined p(J, C, O, I, E, Q). After that, we compute p(J = 1 | C, O, I , E = 1, Q = 1) to answer above question.

Note, the set of all parents of a hidden node is the complete set of mutually exclusive hidden variables and the set of all evidence nodes which are children of a hidden node is the complete set of mutually exclusive evidence variables.

4 Conclusion

There is no doubt that the combination of Bayesian network and overlay model gives us the appropriate approach for user modeling but it has two disadvantages:

- The expense of data storage is high. A Bayesian network which has *n* variables together *n* CPTs with 2^n parameters (values in CPTs) under constraint: "each variable is binary (0 and 1)". If variables are not binary, the number of parameters are huge, so, it is difficult to store them in memory.
- The computation of posterior probability which is basis of inference consumes much time when executing in runtime because it is rather complex.

The first cause which is the inherent attribute of Bayesian network can be only restricted by programming technique when implementing network and it would be best to declare binary variables. On the other hand, it is the done to use CPT instead of continuous probability density/distribution function for solving the second problem.

As already discussed, the structure and parameters (CPTs) in our model are fixed and specified by experts. However, they must be evolved after each occurrence of

evidence. When learning machine is concerned, structural learning is process of gradual improving the structure of model and correspondingly, parametric learning is process of changing the parameters so as to be more suitable. We will discuss the improvement on qualitative model (structure) and quantitative model (parameters) in the other paper.

References

1. Akiba, T., Tanaka, H.: A Bayesian Approach for User Modeling in Dialogue Systems. In: COLING 1994. The 15th International Conference on Computational Linguistics, vol. 1 (1994)
2. Allen, R.B.: User Models: Theory, Method, and Practice. International Journal of Man-Machine Studies 32, 511–543 (1990)
3. Brachman, R.J., Levesque, H.J.: Knowledge Representation and Reasoning. ©2003 CMPT 411/882 Course Home Page (Spring 2005)
4. Bunt, A., Conati, C.: Probabilistic Student Modelling to Improve Exploratory Behaviour. Journal of User Modeling and User-Adapted Interaction 13(3), 269–309 (2003)
5. Conati, C.: Probabilistic Assessment of User's Emotions in Educational Games. Journal of Applied Artificial Intelligence, special issue on " Merging Cognition and Affect in HCI" 16(7-8), 555–575 (2002)
6. Charniak, E.: Bayesian Network without Tears. AI magazine (1991)
7. Chin, D.N.: A Case Study of Knowledge Representation in UC. In: Proceedings of the Eighth International Joint Conference on Artificial Intelligence, August 1983, vol. 1, pp. 388–390. Karlsruhe, West Germany (1983)
8. Chin, D.N.: KNOME: Modeling What the User Knows in UC. In: Kobsa, A., Wahlster, W. (eds.) User Models in Dialog Systems, pp. 74–107. Springer, Heidelberg
9. Fagin, R., Halpern, J.Y.: Reasoning about knowledge and probability. ACM 41(2), 340–367 (1994); Preliminary version appeared in: Vardi, M.Y. (ed.): Second Conf. on Theoretical Aspects of Reasoning about Knowledge, pp. 277-293. Morgan Kaufmann (1988); Corrigendum: J. ACM 45(1), p. 214 (January 1998)
10. Finin, T., Drager, D.: GUMS: A General User Modeling System. In: Proceedings of the Canadian Society for Computational Studies of Intelligence 1986 (CSCSI 1986) (1986)
11. Halpern, J.Y.: Reasoning about Knowledge,
    http://www.cs.cornell.edu/home/halpern/topics.html#rak
12. Halpern, J.Y.: Reasoning about Uncertainty,
    http://www.cs.cornell.edu/home/halpern/topics.html#rau
13. Henze, N., Nejdl, W.: Bayesian Modeling for Adaptive Hypermedia Systems. Knowledge Based Systems Group, University of Hannover, Lange Laube 3, 30159 Hannover, Germany (1999)
14. Horvitz, E., Breese, J., Heckerman, D., Hovel, D., Rommelse, K.: The Lumière Project: Bayesian User Modeling for Inferring the Goals and Needs of Software Users. In: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, Madison, WI, July 1998, pp. 256–265. Morgan Kaufmann, San Francisco (1998)
15. Jameson, A.: Logic is not enough: Why reasoning about another person's beliefs is reasoning under uncertainty. In: Laux, A., Wansing, H. (eds.) Knowledge and Belief in Philosophy and Artificial Intelligence. Akademie Verlag, Berlin (1995)
16. Jameson, A., Hoeppner, W., Wahlster, W.: The Natural Language System HAM-RPM as a Hotel Manager: Some Representational Prerequisites. In: Wilhelm, R. (ed.) GI - 10. Jahrestagung Saarbrücken. Springer, Berlin (1980)

17. Jedlitschka, A., Althoff, K.-D.: Using Case-Based Reasoning for User Modeling in an Experience Management System. In: Proc. Workshop Adaptivität und Benutzermodellierung in Interaktiven Systemen (ABIS 2001), GI-Workshop-Woche Lernen - Lehren - Wissen - Adaptivität (LLWA 2001), Universität Dortmund, October 8-12 (2001)
18. Johansson, P.: User Modeling in Dialog Systems. St. Anna Report: SAR 02-2
19. Kobsa, A.: First experiences with the SB-ONE knowledge representation workbench in natural-language applications. ACM SIGART Bulletin 2(3), 70–76 (1991)
20. Kobsa, A.: User Modeling: Recent Work, Prospects and Hazards. Department of Computer Science, Columbia University, New York, USA (1993)
21. Orwant, J.L.: Doppelgänger Goes To School: Machine Learning for User Modeling. Master's thesis, Massachusetts Institute of Technology (1993)
22. Paiva, A., Self, J.: A Learner Model Reason Maintenance System. In: Cohn, A. (ed.) ECAI 1994. 11th European Conference on Artificial Intelligence. John Wiley & Sons, Ltd., Chichester (1994)
23. Papatheodorou, C.: Machine Learning in User Modeling. In: Paliouras, G., Karkaletsis, V., Spyropoulos, C.D. (eds.) ACAI 1999. LNCS (LNAI), vol. 2049, pp. 286–294. Springer, Heidelberg (2001)
24. Pohl, W.: Logic-Based Representation and Reasoning for User Modeling Shell Systems. In: User Modeling and User-Adapted Interaction (UMUAI 1999), vol. 9(3), pp. 217–283 (1999)
25. Pohl, W., Höhle, J.: Mechanisms for flexible representation and use of knowledge in user modeling shell systems. In: Jameson, A., Paris, C., Tasso, C. (eds.) User Modeling: Proceedings of the Sixth International Conference, UM 1997, CISM 1997. Springer, Vienna (1997)
26. Poole, D.: A Logical Framework for Default Reasoning. Logic Programming and Artificial Intelligence Group, Department of Computer Science, University of Waterloo, Waterloo, Ontario, Canada, N2L3G1 (519), 888–4443 (1987)
27. Rich, E.: User Modeling via Stereotypes. Cognitive Science 3, 329–354 (1979)
28. Sparacino, F.: Sto(ry)chastics: a Bayesian Network Architecture for User Modeling and Computational Storytelling for Interactive Spaces. LNCS. Springer, Heidelberg (2003)
29. Ting, C.Y., Phon-Amnuaisuk, S., Chong, Y.K.: Modeling and Intervening Across Time in Scientific Inquiry Exploratory Learning Environment. Journal of Educational Technology & Society 11(3), 239–258 (2008)
30. Ting, C.Y., Reza Beik Zadeh, M.: Assessing Learner's Scientific Inquiry Skills Across Time: A Dynamic Bayesian Network Approach. In: Conati, C., McCoy, K., Paliouras, G. (eds.) UM 2007. LNCS, vol. 4511, pp. 207–216. Springer, Heidelberg (2007)
31. Tedesco, R., Dolog, P., Nejdl, W., Allert, H.: Distributed Bayesian Networks for User Modeling. In: ELEARN 2006: World Conference on E-Learning in Corporate, Government, Health Care, and Higher Education (2006)
32. wikipedia. Default Logic (2007),
    http://en.wikipedia.org/wiki/Default_logic
33. Wilensky, R.: Some Problems and Proposals for Knowledge Representation. Computer Science Division, University of California, Berkely, Report No. UCB/CSD 87/351 (1987)