Prepublication draft of paper published in
Proceedings of the Fourth International Conference on Intelligent Tutoring Systems,
pp 384-393, Springer-Verlag, 1998.

Data ageing: a technique for discounting old data during student modelling

Geoffrey I. Webb
Mark Kuzmycz
School of Computing and Mathematics
Deakin University
Geelong, Vic, 3217, Australia.

Abstract

Student modelling systems must operate in an environment in which a student's mastery of a subject matter is likely to change as a lesson progresses. A student model is formed from evaluation of evidence about the student's mastery of the domain. However, given that such mastery will change, older evidence is likely to be less valuable than recent evidence. Data ageing addresses this issue by discounting the value of older evidence. This paper provides a formal evaluation of the effects of data ageing. While it is demonstrated that data ageing can result in statistically significant increases in both the number and accuracy of predictions that a modelling system makes, it is also demonstrated that the reverse can be true. Further, the effects experienced are of only small magnitude. It is argued that these results demonstrate some potential for data ageing as a general strategy, but do not warrant employing data ageing in its current form.

1 Introduction

A student modelling system seeks to develop a model of the student from observations of the student's past performance. A fundamental difficulty confronting such a system is that while it seeks to construct a model of the student's current state, the evidence upon which the model is constructed is based upon observations of performance arising from previous states. Education aims to transform the student, from whatever initial state in which the student begins the educational interaction, to a state of mastery of the instructional subject matter. Thus, unless a student starts in a state of mastery, in which case educational interactions are not necessary, successful educational interactions will change student states. In consequence, in a successful educational environment, one should not expect the evidence on which a model is based to reflect the current state of the student. But, such evidence is the primary information available to a system on which to base a model.

Most student modelling systems have ignored this problem, or have circumvented it by avoiding the use of historical data (basing a model on only a single recent action).

Feature Based Modelling has sought to tackle the problem directly, using a mechanism called *data ageing*. Data ageing discounts older evidence, placing greater weight on recent evidence. This is motivated by the assumption that the more time that has elapsed between an observation and the formation of a model, the greater the probability that the evidence is no longer relevant.

The original formulation of this mechanism (Webb & Kuzmycz, 1996) initially assigned each observation a weight of 1. The weight of each observation was then discounted by a set proportion each time another relevant observation was incorporated into the model. The induction system took account of weights when developing a model.

While the system incorporating this mechanism demonstrated high accuracy in the challenging task of predicting elementary subtraction results, no evaluation of the contribution of data ageing as opposed to other aspects of the mechanism was undertaken. This paper rectifies this situation by evaluating the effect of differing levels of data ageing upon system performance.

2 Feature Based Modelling

Feature Based Modelling constructs a black-box model of an agent. It seeks to capture the relationships that hold between the inputs and outputs of the agent, but not the mechanisms that underlie those relationships.

Feature Based Modelling employs a simple form of attribute-value machine learning. The context of an action is described by a set of attribute values called *context features*. Each action is described by a set of attribute values called *action features*.

A model takes the form of a set of *associations*. Each association can be thought of as a production rule associating a set of context features to a single action feature.

A model consists of all associations that satisfy the following conditions:

1. $\#(C- > a) \geq min\_evidence$;
2. $\frac{\#(C->a)}{\#(C->a)+\#(C-> \ a)} \geq min\_accuracy$; and
3. there is no association between a specialisation of C and a sibling of a.

where

- $\#(C- > a)$ is the number of observed cases in which all features in C and feature a have been present;
- $\#(C- > \ a)$ is the number of observed cases in which all features in C and a sibling of a have been present; and
- $min\_evidence$ and $min\_accuracy$ are implementation dependent parameters.

Associations are allowed that are contradicted by some of the evidence in order to accommodate noise, inconsistent behaviour, and changes in behaviour.

Most implementations of FBM have used $min\_evidence$ set to 3 and $min\_accuracy$ set to 0.8. Although $min\_accuracy$ of 0.8 ostensibly allows for an association to be accepted when almost 20% of the evidence contradicts it, clause 3 limits the probability of this occurring by suppressing an association if there is a regularity detected in the contrary evidence.

A set of associations can be used to make predictions. To predict an agent's actions in a particular context it is necessary only to extract the set of associations that have all their context features satisfied by the given context. The set of action features for these associations can then be used to make predictions. The action features might fully specify a precise action, or may simply indicate constraints on possible actions, depending both upon the types of action features that are employed and the number of associations that match the current context.

The interested reader is referred to Webb and Kuzmycz (1996) for a more detailed description of Feature Based Modelling and its application.

3 Between and within test data ageing

Feature Based Modeling was initially developed in the context of an intelligent tutoring system (Webb, 1991). In this context, the evidence from a student was presented to the system in the order in which it was generated by the student (each student response was observed immediately). Further, each task followed immediately one after the other. In such a setting it seemed sensible to discount data at a single set rate after each interaction with the student.

However, this design decision may be less appropriate in other domains. Recent research has considered the application of Feature Based Modelling to the development of models of subtraction performance in a non-tutoring environment (Webb & Kuzmycz, 1996). Students are administered a sequence of tests, each comprising 40 three column subtraction problems. Tests are administered at one week intervals. After each test, individual student models are created from the test results for the student to date. Each such model is used to predict the student's answers to the questions on the next test. It is not possible for the computer to determine the order in which questions are answered during a test. However, the original data ageing mechanism caused the results for the first question processed by the computer to be discounted 117 times (each column is treated as a separate task, so three discountings occur for each subsequent three column problem that is examined). In contrast, the data for the last column of the last problem will not be discounted at all. This does not appear to be appropriate as:

- a student may tackle the problems in a different order to that in which the system processes them; and
- the changes in the student's state over the interval between tests is likely to be greater than that between tasks within a single test, but greater discounting occurs during a test than between tests.

To address these issues it would appear appropriate to perform discounting between tests rather than discounting within tests.

In more abstract terms, what is being advocated is a break in the link between individual tasks and the data ageing schedule. Rather than discounting the data after each task, it should be possible to specify a data ageing schedule that reflects the probable impact on the relevance of data within the model of different intervals between observations. One can imagine a schedule that is sensitive to whether specific forms of educational interaction have occurred (lectures, practical work, etc.) and the interval of time between observations.

4 Evaluation

Previous evaluation of Feature Based Modelling has each time employed a set discounting rate (Kuzmycz & Webb, 1992; Webb & Kuzmycz, 1996). This has not permitted the evaluation of any of:

- the relative performance of different discounting rates; or
- whether discounting is in itself beneficial.

To evaluate these factors Feature Based Modelling was applied to a large modelling task using discounting rates of 0%, 5%, 10%, 20% and 50%.

The task to which each of these alternatives was applied was the task used by Webb and Kuzmycz (1996) and Webb, Chiu and Kuzmycz (1996). This involves a body of results of three column subtraction tests administered to eight to nine year old Australian primary school students.

This data was collected as follows:

1. 73 nine to ten year old primary school students were divided into two treatments: Random and Error Repeat. This was achieved by sorting the students at each of the three participating schools into alphabetical order and then assigning them to alternating treatments in order.

2. An initial set of 40 three column subtraction problems were randomly generated as follows:

$$minuend = (random() \text{ modulo } 900) + 100$$

$$subtrahend = random() \text{ modulo } (minuend + 1)$$

   where $random()$ is a pseudo random number generator that generates 32 bit unsigned integer values. This resulted in random three digit subtraction problems such the minuend contained three digits and the correct result was positive.

3. The initial test was presented to all subjects.

4. The following was repeated three times

   (a) a new set of 40 three column subtraction problems were generated as follows:

      i. for each subject in the Error Repeat treatment, all problems from the last test sheet for which the subject made an error were copied to the new problem set and then new random problems were generated, as per step 2, to make a total of 40 problems,

      ii. for each subject in the Random treatment, 40 random problems were generated, as per step 2.

   (b) the tests were administered to the subjects.

5. A final single set of 40 three column subtraction problems was generated as per step 2.

6. The final test was administered to all subjects.

Successive tests were all administered at weekly intervals. (As variations over time were not relevant, Webb and Kuzmycz (1996) used only the first two of this series of tests.)

Normal tuition proceeded between tests. Thus, students' approaches to the domain could be expected to alter between sessions.

The following evaluation is performed by forming models from a sequence of tests, 1, ... n, which is then applied to predict the subject's precise answers in test $n + 1$. This process is repeated for n set to each of $2, 3$ and 4.

The same modelling and prediction techniques were employed as in Webb and Kuzmycz (1996) except that the data ageing procedures were systematically manipulated.

The following are lists of the context and action features employed to model subtraction skills in this study. For more detail, the interested reader is directed to Webb and Kuzmycz (1996).

Context Features: Minuend $>$ Subtrahend; Minuend $<$ Subtrahend; Minuend = Subtrahend; Minuend $>$ Subtrahend in the column to the right; Minuend $<$ Subtrahend in the column to the right; Minuend = Subtrahend in the column to the right; Minuend $>$ Subtrahend two columns to the right; Minuend $<$ Subtrahend two columns to the right; Minuend = Subtrahend two columns to the right; Minuend is zero; Minuend is not zero; Minuend is zero in the column to the left; Minuend is not zero in the column to the left; Minuend is zero in the column to the right; Minuend is not zero in the column to the right; Minuend is one in the column to the left; Minuend is not one in the column to the left; Subtrahend is zero; Subtrahend is not zero; Subtrahend is nine; Subtrahend is not nine; Subtrahend is nine in the column to the right; Subtrahend is not nine in the column to the right; Subtrahend is blank; Subtrahend is not blank; This column is right-most; This column is left-most; This column is neither left nor right-most.

Action Features: Result = Minuend − Subtrahend; Result = Minuend − Subtrahend − 1; Result = Minuend − Subtrahend + 10; Result = Minuend − Subtrahend + 9; Result = Minuend − Subtrahend + 8; Result = Minuend; Result = Subtrahend; Result = zero; Result = Minuend − Subtrahend − 2; Result = Subtrahend − Minuend; Result is correct; Result is incorrect.

Each prediction relates to a single column of a subtraction problem. The prediction is considered correct if the precise digit for the column is predicted. For some columns no predictions will be made. This can arise for a number of reasons:

1. no associations apply to the column;
2. multiple associations apply that make differing predictions; or
3. the associations that do apply do not fully specify a single digit.

5 Results

The treatments were evaluated on six metrics:

Cover: the percentage of columns for which a prediction was made;

Accuracy: the percentage of predictions that were correct;

Student error cover: the percentage of those columns for which a student made an error, for which a prediction was made;

Student error accuracy: the percentage of the student error cover for which the predictions were correct;

Error predictions: the number of predictions that a student would make an error; and

Error prediction accuracy: the percentage of error predictions that were correct.

Figures 1 to 6 summarise the round by round performance on these metrics. Each of these figures plots the relevant outcome for each condition. Each of the data ageing levels are labelled by the percentage level.

All treatments are equivalent for the first set of predictions (round 2), as ageing of data only occurs immediately before the round 2 data is incorporated into the model and thus has no effect at the time when predictions are made with respect to round 2.

It can be seen that moderate levels of data ageing lead to small increases in both the numbers (Figure 1) and accuracy (Figure 2) of predictions for rounds 3 and 4.

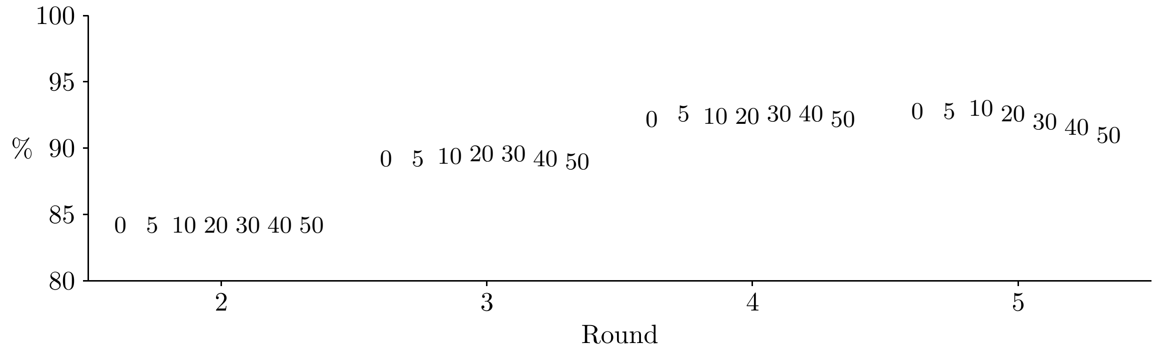

Figure 1: Round by round cover

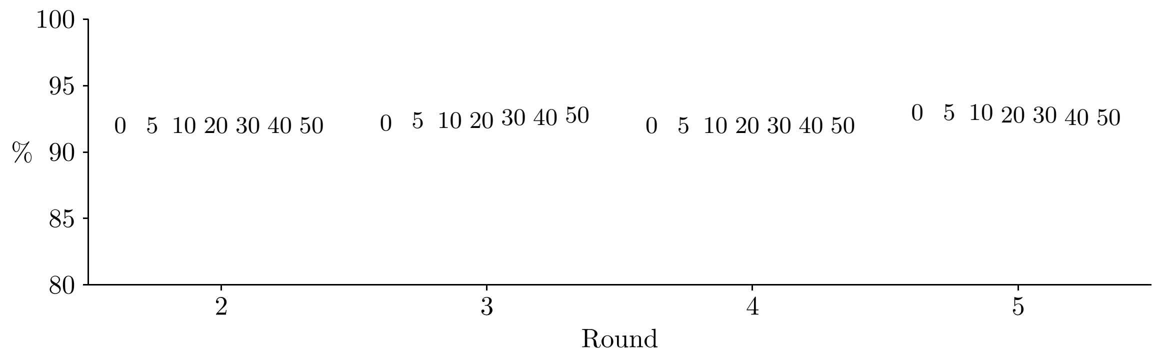

Figure 2: Round by round accuracy

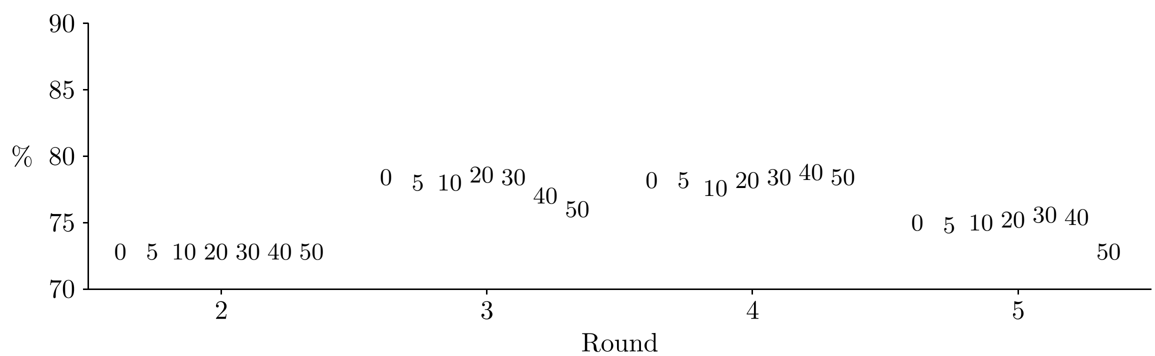

Figure 3: Round by round student error cover

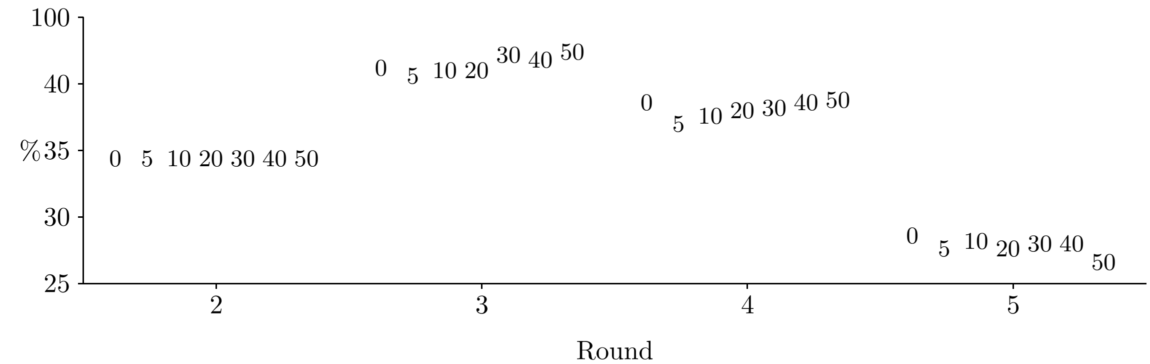

Figure 4: Round by round student error accuracy

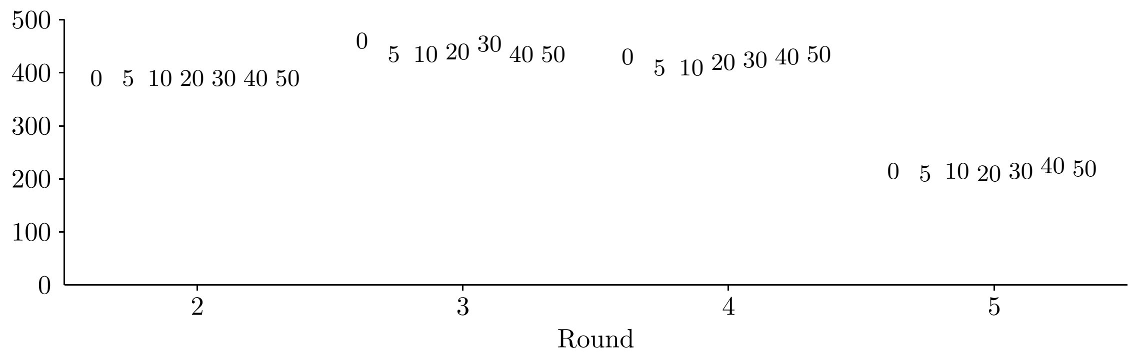

Figure 5: Round by round error predictions

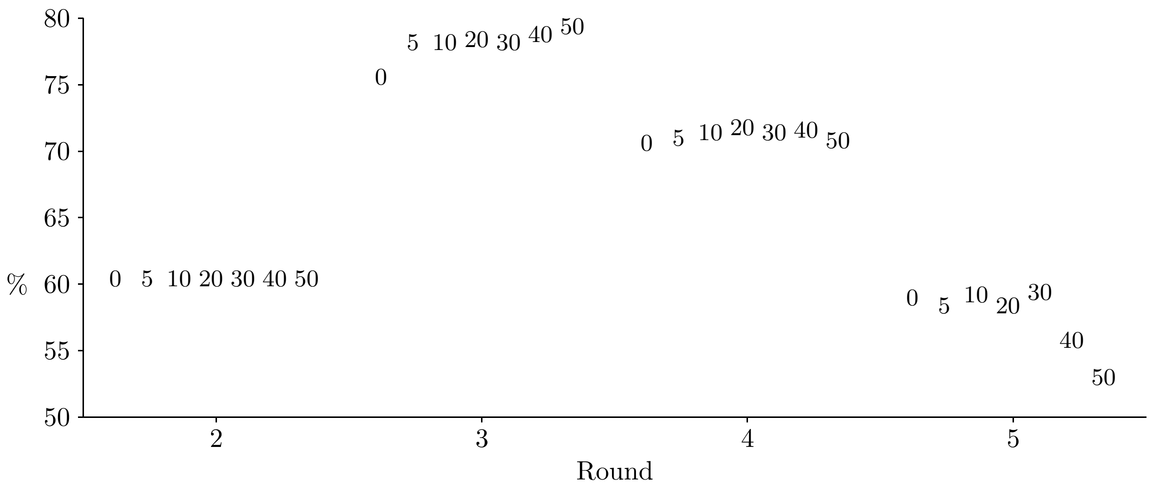

Figure 6: Round by round error prediction accuracy

Criterion	Round 3			Round 4			Round 5		
	0%	30%	p	0%	30%	p	0%	30%	p
Cover	39	78	0.000	57	52	0.351	60	29	0.001
Accuracy	36	64	0.003	43	50	0.267	47	30	0.034
Student error cover	?	?	?	?	?	?	?	?	?
Student error accuracy	15	18	0.364	16	22	0.209	14	10	0.271
Error predictions	20	15	0.249	20	17	0.371	17	10	0.124
Error prediction accuracy	13	12	0.500	9	15	0.154	13	9	0.262

Table 1: Sign tests comparing 0% and 30% data ageing

However, 20% and higher ageing values lead to small decreases in both numbers and accuracy of predictions on the last round. There is no clearly discernible pattern to how ageing affects the number of errors for which a prediction is made (Figure 3), the accuracy of those predictions (Figure 4) or the number of predictions that the student would make an error (Figure 5). However, data ageing does appear to improve the accuracy of error predictions (Figure 6) for rounds 3 and 4, although differing levels of ageing lead to greatly differing outcomes on round 5.

Visual inspection of these results figures suggests that data ageing at 30% leads to reasonable overall performance. A statistical comparison of the performance of this level of data ageing against no data ageing was performed. For each round, the relative performance of no data ageing and 30% data ageing was compared on a subject by subject basis. On each of the six criteria examined in the figures, the number of subjects for which each treatment outperformed the other was determined. A binomial sign test was used to evaluate whether there was a statistically significant difference in the performance of the two treatments on the criterion. The results of this evaluation are presented in Table 1. (The columns labelled 0% provide counts of the number of times no data ageing outperformed 30% data ageing, and those labelled 30% provide counts of the number of times 30% data ageing outperformed no data ageing.)

Table 1 shows that 30% data ageing is leading to increases in cover and accuracy significantly more often than not for round 3 but that such an advantage is not apparent for round 4 and the reverse is true for round 5.

With respect to student error cover and student error accuracy, there is no significant advantage to either treatment in any of the rounds.

With respect to the predictions that the student would make an error, neither treatment demonstrated a significant advantage with respect to either the number or accuracy of such predictions.

6 Discussion

The significant increases in both cover and accuracy on round 3 demonstrates that data ageing can provide a significant benefit. The significant losses incurred on round 5 demonstrate that such benefits are far from guaranteed, however.

There is some reason to believe that round 5 might be atypical in some respects. It should be recalled that all subjects were given a single randomly generated test on this round. All treatments experience a significant drop in average performance on all metrics for this round. This suggests that the single test presented to the students was not typical.

That significant improvements on round three are only reflected in the overall performance of the system rather than in the performance in predicting errors could reflect either of two things. It might indicate that discounting some of the student's errors enables the system to more successfully model those aspect of the domain that a student has mastered. On the other hand, it could be that the numbers of errors are so low that the power of the statistics employed is not great enough to uncover such advantage that might exist.

7 Conclusions and further research

A student's mastery of a domain can be expected to change over time. If student modelling is to fulfill its promise, it must allow for such changes. Basing models on a single most recent action will not suffice as, in many cases, a single action will not provide enough information to permit accurate diagnosis. Data ageing seeks to overcome this problem by discounting older evidence in favour of more recent evidence.

However, a formal evaluation of data ageing has led to inconclusive results. 30% data ageing demonstrates a small but statistically significant improvement in the number and accuracy of predictions on the third of a series of five tests, but leads to small but statistically significant decreases in performance on the fifth test. While there is some reason to doubt the typicality of the fifth test, it is nonetheless clear that data ageing in its current form is not providing large improvements in predictive accuracy.

The small but significant improvements in performance that are demonstrated in the third round suggest that there is some merit in the general strategy. To this end it might be valuable to explore alternative approaches to data ageing. One option that might be worth exploring is differential ageing rates. One possibility is that errors should be discounted more rapidly than correct performance, given the evidence that students only repeat an error on approximately one third of occasions (Webb & Kuzmycz, 1996). Another possibility is that data that has been aged once should subsequently be aged at a lower rate, on the assumption that there is less difference between the relevance of evidence that is two or three weeks old than between that of evidence that is one or two weeks old.

In summary, data ageing in its current form does not appear to warrant use. Experimental evaluation suggests, however, that there may be merit in further refinement of the technique.

References

Kuzmycz, M., & Webb, G. I. (1992). Evaluation of feature based modelling in subtraction. In Frasson, C., Gauthier, G., & McCalla, G. I. (Eds.), *Intelligent Tutoring Systems*, pp. 269–276 Berlin. Springer-Verlag.

Webb, G. I. (1991). Inside the unification tutor: The architecture of an intelligent educational system. In Godfrey, B. (Ed.), *Proceedings of the Eighth Annual Conference of the Ausralian Society for Computers in Learning in Tertiary Education*, pp. 675–684 Launceston. ASCILITE.

Webb, G. I., Chiu, P., & Kuzmycz, M. (1996). A comparative evaluation of the use of C4.5 and Feature Based Modelling as induction engines for input/output agent modelling. *Submitted for publication*.

Webb, G. I., & Kuzmycz, M. (1996). Feature Based Modelling: A methodology for producing coherent, consistent, dynamically changing models of agent's competencies. *User Modeling and User Assisted Interaction*, *5*(2), 117–150.