Knauf, Rainer; Tsuruta, Setsuo; Gonzalez, Avelino J.:

Towards modeling human expertise: an empirical case study

Zuerst erschienen in:
Proceedings of the eigtheenth International Florida Artificial Intelligence Research Society conference : [Clearwater Beach, Florida, May 15 - 17, 2005] / ed. by Ingrid Russel ....
Menlo Park, Calif : AAAI Press, 2005
Referenz-Link AAAI:
http://www.aaai.org/Library/Conferences/FLAIRS/FLAIRS-2005/Abstracts/flairs05-039.html

Towards Modeling Human Expertise: An Empirical Case Study

Rainer Knauf
Faculty of Computer Science and Automation
Ilmenau Technical University
PO Box 100565, 98684 Ilmenau, Germany

Setsuo Tsuruta
School of Information Environment
Tokyo Denki Unversity
2-1200 Musai-gakuendai, Inzai
Chiba, 270-1382, Japan

Avelino J. Gonzalez
Dept. of Electrical and Computer Engineering
University of Central Florida
Orlando, FL 32816-2450, USA

Abstract

The success of TURING Test technologies for system validation depends on the quality of the human expertise behind the system. The authors developed models of collective and individual human expertise, which are shortly outlined here. The focus of the paper is an experimental work aimed at determining the quality of these models. The models have been used for both solving problem cases and rating (other agents') solutions to these cases. By comparing the models' solutions and ratings with those of the human original we derived assessments of their quality. An analysis revealed both the general usefulness and some particular weaknesses.

Introduction

To make TURING TEST validation results less dependent on the experts' opinions and to decrease the workload of the experts, a Validation Knowledge Base (*VKB*) was developed as a model of collective human expertise of former expert panels and Validation Expert Software Agents (*VESA*) were developed as a model of individual human expertise (Tsuruta et.al. 2002; Knauf et al. 2004a; Knauf et al. 2004c). These concepts have been implemented in a validation framework (Knauf et al. 2002). To estimate the usefulness of these concepts and to reveal their weaknesses, a prototype test was performed (Knauf et al. 2004b). The purpose of the present paper is to report the basic insights about the use of both models, *VKB* and *VESA*, in an experimentation environment.

The paper is organized as follows: The next section provides a short summary about the concepts developed so far: the validation framework, *VKB* and *VESA*. Section three describes the prototype application scenario. In section four, main results are presented and concept improvements are derived. The fifth section summarizes the paper.

The Concepts so far

The TURING Test validation framework covers five steps: (1) test case generation, (2) test case experimentation, (3) evaluation of results, (4) validity assessment, and (5) system refinement (Knauf et al. 2002). The most expensive step is the 2^{nd} one, because of the necessary human involvement. This step is supported by a *VKB* (Tsuruta et.al. 2002; Knauf et al. 2004c), which contains validation knowledge of previous validation processes. Validation knowledge, in this context, is a set of former test cases with their most accepted (best rated) solutions. Furthermore, a *VESA* has been developed (Tsuruta et.al. 2002; Knauf et al. 2004c) to keep validation knowledge, such as previous validation judgments or the experiences of human experts. It is an intelligent agent corresponding to a particular human. *VESA*s systematically model human validators by keeping the personal validation knowledge of their corresponding experts and analyzing similarities with other experts. At some point, a *VESA* may be able to serve as a temporary substitute for a missing human expert.

The *VKB* is a database of test cases and their associated solutions that received an optimal rating in previous validation sessions. The information stored and maintained in the *VKB* for use in the test case experimentation consists of the required input data, the produced output data, and some necessary additional information. According to the formal settings in (Knauf et al. 2002) and (Kurbad 2003), the *VKB* contains a set of previous (historical) test cases, which can be described by 8-tuples $[t_j, E_K, E_I, sol^{opt}_{Kj}, r_{IjK}, c_{IjK}, \tau_S, D_C]$, where t_j is a test data (a test case input), sol^{opt}_{Kj} is a solution associated to t_j, which gained the maximum experts' approval in a validation session, E_K is a list of experts who provided this particular solution, E_I is a list of experts who rated this solution, r_{IjK} is the rating of this solution, which is provided by the experts in E_I, c_{IjK} is the certainty of this rating, τ_S is a time stamp associated with the validation session in which the rating was provided, and D_C is an informal description of the application domain C that is helpful to explain similarities between different domains or fields of knowledge. Additionally, a list of supporters $E_S \subseteq E_I$ for each solution sol^{opt}_{Kj} is kept in *VKB*. A supporter is a rating expert who provided a positive rating for sol^{opt}_{Kj}.

For example, a part of *VKB* in the prototype test (as described in section 3) looks like shown in table 1. Here, e_1, e_2, and e_3 are particular (real) human experts, $o_1, .., o_{25}$ are test case outputs (solutions), and the time stamps are represented by natural numbers $1, .., 4$. The *VKB* is built within the first validation session, in which *all* test case inputs along with their optimal solutions are the subject of a new entry. It

t_j	E_K	E_I	sol_{Kj}^{opt}	r_{ijk}	c_{ijk}	τ_S	D_C
t_1	e_1, e_3	$[e_1, e_2, e_3]$	o_6	$[1, 0, 1]$	$[0, 1, 1]$	1	
t_1	e_2	$[e_1, e_2, e_3]$	o_{17}	$[0, 1, 0]$	$[1, 1, 1]$	4	
t_2	e_1, e_3	$[e_1, e_2, e_3]$	o_7	$[0, 0, 1]$	$[0, 0, 1]$	1	
...	...	...	...	...	...	...	...

Table 1: An example for *VKB*'s entries

is updated in the following sessions by adding all examined test cases of this session. There is no "updating" of existing entries. This because at least the time stamp differs from the ones of the existing entries, i.e. validation knowledge gained at different sessions from different entries.

VKB functions in the second step, the test case experimentation. In the original approach, the test case generation procedure consist of two steps (a) generating a quasi exhaustive set of test cases *QuEST* and (b) reducing it down to a reasonably sized set of test cases *ReST*(Knauf et al. 2002). Exactly between these two sub–steps is the "entry–point" of the external validation knowledge stored in a *VKB* that has been constructed in prior validation sessions. Both *QuEST* and the historical cases in *VKB* are subjected to the criteria–based reduction procedure that aims to build a subset of test cases in *QuEST* or *VKB*. The cases in *VKB* are included in the reduction process to (1) ensure that they meet the requirements of the current application and (2) their number is small enough to be the subject of the time consuming and expensive test case experimentation. The *VKB*, therefore is a database of test cases and their associated solutions that received an optimal rating in previous validation sessions. These solutions are considered an additional (external) source of expertise that did not explicitly appear in the solving session, but it is a subject of the rating session. Regardless of their former ratings, the cases originated from the *VKB* have to be rated by the current expert panel in the current session for the reasons explained (Knauf et al. 2004c).

A *VESA* is requested, in case an expert e_i is not available to solve a case t_j. e_i's former (latest) solution is considered by this expert's *VESA*. It is assumed that e_i still has the same opinion about t_j's solution. Thus, *VESA* provides this solution. If e_i never considered case t_j before, similarities with other experts who might have the same "school" or "thinking structures" are considered. Among all experts who ever provided a solution to t_j , the one with the largest subset of the solutions like e_i's for the other cases that both solved is identified as the one with the most similar behavior. e_i's solution is assumed to be the same as this other expert's. This solution is consequently adopted by the *VESA* that corresponds to the missing expert. Formally, a *VESA*$_i$ acts as follows when requested to provide an assumed solution of expert e_i for a test case input t_j:

1. In case e_i solved t_j in a former session, his/her solution with the latest time stamp will be provided by *VESA*$_i$.
2. Otherwise,
   (a) All validators e', who ever delivered a solution to t_j form a set $Solver_i^0$, which is an initial dynamic agent for e_i: $Solver_i^0 := \{e' : [t_j, E_K, ...] \in$ *VKB* $, e' \in E_K\}$
   (b) Select the most similar expert e_{sim} with the largest set of cases that have been solved by both e_i and e_{sim} with the same solution and in the same session. e_{sim} forms a refined dynamic agent $Solver_i^1$ for e_i: $Solver_i^1 := e_{sim} : e_{sim} \in Solver_i^0, |\{[t_j, E_K, \_, sol_{Kj}^{opt}, \_ , \_ , \tau_S, \_] : e_i \in E_K, e_{sim} \in E_K\}| \rightarrow max!$
   (c) Provide the latest solution of the expert e_{sim} to the present test case input t_j , i.e. the solution with the latest time stamp τ_S by *VESA*$_i$.
3. If there is no such most similar expert, provide $sol := unknown$ by *VESA*$_i$.

If a *VESA*$_i$ is requested to provide assumed rating of expert e_i to a solution of a test case input t_j, it models the rating behavior of e_i as follows:

1. If e_i rated t_j before, look at the rating with the latest time stamp τ_S, *VESA*$_i$ provides the same rating r and the same certainty c on behalf of e_i.
2. Otherwise,
   (a) All validators e' , who ever delivered a rating to t_j form a set $Rater_i^0$, which is an initial dynamic agent for e_i: $Rater_i^0 := \{e' : [t_j, \_ , E_I, ...] \in$ *VKB* $, e' \in E_I\}$
   (b) Select the most similar expert e_{sim} with the largest set of cases that have been rated by both e_i and e_{sim} with the same rating r and in the same session. e_{sim} forms a refined dynamic agent $Rater_i^1$ for e_i : $Rater_i^1 := e_{sim} : e_{sim} \in Rater_i^0, |\{[t_j, \_ , E_I, sol_{Kj}^{opt}, r_{IjK}, \_ , \tau_S, \_] : e_i \in E_I, e_{sim} \in E_I, \}| \rightarrow max!$
   (c) Provide the latest rating r along with its certainty c to t_j of e_{sim} by *VESA*$_i$.
3. If there is no most similar expert e_{sim}, provide $r := norating$ along with a certainty $c := 0$ by *VESA*$_i$.

Table 2 shows an example that indicates a *VESA*'s behavior in a solution session that took place within the prototype experiment (see section 3). The experiment was intended to compare a *VESA*'s behavior (*VESA*$_2$, in the example) with the behavior of its human counterpart (e_2, in the example) to validate the *VESA* approach. t_i are test case inputs and o_i are the outputs provided by the *VESA* respectively the associated human expert. EK_3 denotes the "external knowledge" of the *VKB* within the 3rd session, i.e. test cases with

EK_3	solution of		EK_3	solution of	
	VESA$_2$	e_2		*VESA*$_2$	e_2
t_{29}	o_8	o_8	t_{36}	o_9	o_9
t_{30}	o_9	o_9	t_{37}	o_9	o_9
t_{31}	o_2	o_2	t_{38}	o_9	o_9
...	...	...	...	...	...

Table 2: An example for a *VESA*'s solving behavior

inputs, for which there is also an entry in the *VKB*. Here, in only one of the 14 test cases *VESA*$_2$ (the model of the expert e_2) behaved different from its human counterpart.

Table 3 serves as an example that shows a *VESA*'s behavior in a rating session that took place within the prototype experiment. Again, EK_3 denotes the "external knowledge" of the *VKB* within the 3rd session. Possible ratings are 1 ("correct solution to this test case input") and 0 ("incorrect solution to this test case input"). Here, in seven out of the 24 test cases *VESA*$_2$ (the model of the expert e_2) behaved different from its human counterpart.

EK_3	solution	rating of	
		VESA$_2$	e_2
t_1	o_4	0	0
t_1	o_{18}	1	1
t_2	o_{20}	0	1
...	...	...	...

Table 3: An example for a *VESA*'s rating behavior

Actually, to learn a model of the human experts' problem solving behavior, *VESA* still depends on the knowledge of human validators. Learning in the concept of *VESA* is analyzing the solving and rating performance of human experts. The quality of the learning results, i.e. the quality of *VESA*, depends on the quantity and coverage of data provided by the human experts. Therefore, on the one hand, a *VESA* is able to replace its human source temporarily. However, on the other hand, a *VESA* deteriorates if it does not acquire human input over an extended period of time. A concept to check whether or not a *VESA* is still valid is outlined in the refinement section below.

The Prototype Application Scenario

Validation of validation approaches is at least as time consuming as the validation itself. In fact, all the problems with the human resources to perform the evaluation of our approach occur at least to the same degree. *How to find human experts who are able and willing to take part in an experiment without compensating them for their workload?*

One possibility is to choose an application field in which the entertainment factor exceeds the workload factor. Thus, the authors decided to choose an amusing application problem: The selection of an appropriate wine for a given dinner.

By consulting the topical literature, we derived some informal knowledge and developed an intelligent system (in the form of a rule–based system) as a subject of validation.

The Knowledge Base

Basically, the issue of selecting an appropriate wine depends on three inputs the main course s_1, the kind of preparation s_2, and the style of its preparation s_3.

The input space of the considered classification problem is $I = \{[s_1, s_2, s_3]$ with $s_1 \in \{$*pork, beef, fish, ..*$\}$, $\}$, $s_2 \in \{$*boiled, grilled, ..*$\}$, and $s_3 \in \{$*Asian, Western*$\}$. The output $O = \{o_1, \ldots, o_{24}\}$ contains 24 different kinds of wine (Knauf et al. 2004c)[1]:

$o_1 = Red\ wine, fruity, low\ tannin, less\ compound$
$o_2 = Red\ wine, young, rich\ of\ tannin$
...

Expressing the informal knowledge with these input and output specification as HORN clauses leads to a rule base R consisting of 45 rules (Knauf et al. 2004c):

r_1 $o_1 \leftarrow (s_1 = fowl)$
r_2 $o_1 \leftarrow (s_1 = veal)$
r_3 $o_2 \leftarrow (s_1 = pork) \wedge (s_2 = grilled)$
...

The Test Cases

According to the test case generation technique as described in (Knauf et al. 2002), we formally computed a *Quasi Exhaustive Set of Test Cases* (*QuEST*) that contains 145 cases (see (Knauf et al. 2004c) for details of the computation). To generate the *Reasonable Set of Test Cases* (*ReST*), we applied four criteria according to the semantic of the test cases and received 42 test inputs form the reasonable set of test cases *ReST*.

Application Conditions

Available resources were three human experts (e_1, e_2, e_3) and the reasonable set of test cases *ReST*$= t_1, ..., t_{42}$. The desired outcome are answers to the following questions:

1. *Does the VKB contribute to the validation sessions at an increasing rate with an increasing number of validation sessions?* How many external solutions (outside the expertise of the current expert panel) are introduced into the rating process by the *VKB*?
2. *Does the VKB contribute valid knowledge (best rated solutions) in an increasing rate with an increasing number of validation sessions?* How many of the introduced solutions win the rating contest against the solutions of the current expert panel?
3. *Does the VKB increasingly gain the human expertise as number of validation sessions increases?* How many new best rated solutions are introduced into the *VKB* after a validation session?
4. *Do the VESAs model of their human source improve with an increasing number of validation sessions?* Do the *VESA*s provide the same solutions and ratings as their human counterpart?

Each of the three experts as well as the rule base was asked to solve the 42 test cases above in four sessions with 28 test cases each (i.e. some test cases repetitively)[2]. The session plan is shown in table 4. Each session leads to an updated *VKB* as well as to updated *VESA*s for each of the three experts e_1, e_2, and e_3.

- For the *VKB*, every optimal (best rated) solution sol_j^{opt} to a test input t_j (see (Knauf et al. 2002) for details of computing it) is stored in the *VKB* along with (a) a list of experts who provided this solution, (b) a list of experts, who provided ratings (along with their certainties) to this solution and (c) their ratings and certainties, and (d) a time stamp that indicates when the current session was stored.

[1] This is the initial output set. Of course, the human expertise might bring new outputs in the process.

[2] The repetition of cases in later sessions is intended to realize the change of opinions of the experts over time, because the *VESA*s need to follow these changes.

#	experts			*VESA*s			*ReST*
	e_1	e_2	e_3	1	2	3	
1	+	+	+	–	–	–	$ReST^1 = \{t_1, \ldots, t_{28}\}$
2	$\oplus$	+	+	+	–	–	$ReST^2 = \{t_{15}, \ldots, t_{42}\}$
3	+	$\oplus$	+	–	+	–	$ReST^3 = \{t_1, \ldots t_{14}, t_{29}, \ldots, t_{42}\}$
4	+	+	$\oplus$	–	–	+	$ReST^4 = \{t_i : t_i \mod 3 \neq 0\}$

+ takes part – does not take part $\oplus$ takes part for comparing with *VESA*

Table 4: Scheduled Validation Sessions

- For the *VESA*s, which are used in a current session (indicated by "+" in table 4) their behavior (i.e. their provided solutions and ratings) is computed.

We refer to the resulting *VKB*s and *VESA*s[3] of an i–th session as VKB^i, $VESA_1^i$, $VESA_2^i$, and $VESA_3^i$. Again, the one *VKB* contains collective knowledge gained in former sessions while the several *VESA*s model individual knowledge of a particular expert. $ReST^i$, on the other hand, is the set of test cases generated for the current session, i.e. its top index is larger than that of the *VESA*s by one, because their indices refer to the current session whereas the *VKB*'s and *VESA*s' indices refer to the result of the preceding session.

For a fair evaluation of the usefulness of *VKB*, the intersection of test case inputs in *VKB* and *ReST* (EK = external knowledge) needs to be considered in each session, because this is the only knowledge that has a chance to be introduced from outside the current human expertise into the rating process by the *VKB*:[4]

$$\begin{array}{lcll} EK^1 & = & \emptyset \cap ReST^1 & = \emptyset \\ EK^2 & = & \Pi_1(VKB^1) \cap ReST^2 & = \{t_{15}, \ldots, t_{28}\} \\ EK^3 & = & \Pi_1(VKB^2) \cap ReST^3 & = ReST^3 \\ EK^4 & = & \Pi_1(VKB^3) \cap ReST^4 & = ReST^4 \end{array}$$

The cardinalities of these sets are $|EK^1| = 0$, $|EK^2| = 14$, $|EK^3| = |EK^4| = 28$. For the evaluation (see the four questions at the beginning of this section) of the scheduled four sessions, we determine after each session (session # i), beginning with the second session[5]

- the number $rated_i$ of cases from VKB^{i-1}, which were the subject of the rating session and relate it to $|EK^i|$: $Rated_i := rated_i/|EK^i|$
- the number $best_i$ of cases from VKB^{i-1}, which provided the optimal (best rated) solution and relate it to $|EK^i|$: $BestRated_i := best_i/|EK^i|$
- the number $intro_i$ of cases from VKB^{i-1}, for which a new solution has been introduced into *VKB* and relate it to $|EK^i|$: $Introduced_i := intro_i/|EK^i|$
- the number $ident_i$ of solutions and ratings, which are identical responses of e_{i-1} and $VESA^{i-1}$ and relate it to the number of required solutions and ratings: $ModelRating_i := ident_i/|required\_responses|$

[3] $VESA_1$, $VESA_2$, and $VESA_3$, which model the behavior of the experts e_1, e_2, and e_3.

[4] $\Pi_1(VKB^i)$ denotes the 1st projection, i.e. the set of the 1st elements of the 8–tuples in *VKB*. $|EK^i|$ denotes the cardinality of the set EK^i, i.e. the number of its elements.

[5] In the first session the *VKB* is empty and thus, not able to contribute any external knowledge.

The above four questions can now be addressed as follows: (1) $Rated_4 > Rated_3 > Rated_2$, (2) $BestRated_4 > BestRated_3 > BestRated_2$, (3) $Introduced_4 < Introduced_3 < Introduced_2$, and (4) $ModelRating_4 > ModelRating_3 > ModelRating_2$.

Results and Refinements

On the Usefulness of VKB and VESA

Because of the above mentioned problems with the interpretation, the results in terms of the four questions to indicate the benefit of *VKB* and *VESA* (as introduced and quantified above) the step from the 3^{rd} to the 4^{th} session does reflects the truth much better than the step from the 2^{nd} to the 3^{rd} session. The four questions are addressed as follows with respect to the computation of the $Rated_i$, $BestRated_i$, $Introduced_i$, and $ModelRating_i$:

1. $Rated_4 > Rated_3 > Rated_2$?
   - In the 2^{nd} session there was 1 case (out of 14), for which VKB^1 had a solution which was not in the process anyway: $rated_2 = 1$, $Rated_2 := 1/14$. In the 3^{rd} session there were 2 cases for which VKB^2 had a solution which was not in the process anyway: $rated_3 = 2$, $Rated_3 := 2/28$. In the 4^{th} session, there were 24(!) cases, for which VKB^3 had a solution which was not in the process anyway: $rated_4 = 24$, $Rated_4 := 24/28$.
   - With $Rated_4 \approx 0.85$, $Rated_3 \approx 0.071$, and $Rated_2 \approx 0.071$ this requirement was met at least in the step from the 3^{rd} to the 4^{th} session.
   - The contribution effect could not really be expected as a result of the sessions before that. A *VKB* needs to gain a certain amount of "historical experience", before it can contribute to a new session sufficiently. Indeed, after the 3^{rd} session, a remarkable number (24 out of 28) possible cases of VKB^3 have been introduced in the rating process. A 5^{th}, 6^{th} and further sessions would conceivably show this effect much more convincingly.
2. $BestRated_4 > BestRated_3 > BestRated_2$?
   - In the 2^{nd} session, the one solution which was introduced by VKB^1 did not become the optimal one: $best_2 = 0$, $BestRated_2 := 0$. Both of the solutions from VKB^2 introduced in the 3^{rd} session, did not become optimal in the rating process: $best_3 = 0$, $BestRated_3 := 0$. Two of the 24 cases that have been submitted by VKB^3 to the 4^{th} session became the optimal solution: $best_4 = 2$, $BestRated_4 := 2/28$.
   - With $BestRated_4 \approx 0.071$, $BestRated_3 = 0$, and $BestRated_2 = 0$ this requirement was also met when going from the 3^{rd} to the 4^{th} session.

- In the 4^{th} session *VKB*^{3} contributed solutions for two cases, that had not been provided by the human experts, but won the "rating contest". This is the intended effect: The *VKB* introduced new knowledge which turned out to be more valid than the knowledge provided by the human experts.

3. $Introduced_4 < Introduced_3 < Introduced_2$?
   - For $intro_2 = 7$ of 14 cases in EK^2 of the 2^{nd} session, a new solution has been introduced into *VKB*^{1} towards *VKB*^{2}: $Introduced_2 := 7/14$. For $intro_3 = 16$ of 28 cases in EK^3 a new solution has been introduced into *VKB*^{2} towards *VKB*^{3}: $Introduced_3 := 16/28$. For $intro_4 = 17$ of 28 cases in EK^4 a new solution has been introduced into *VKB*^{3} towards *VKB*^{4}: $Introduced_4 := 17/28$.
   - With $Introduced_4 \approx 0.61$, $Introduced_3 \approx 0.57$, and $Introduced_2 = 0.5$ this requirement was not met.
   - The underlying assumption for this question a static domain knowledge, which needs to be explored systematically. However, this was not true for the considered domain. In interesting problem domains there is change over time of both the domain knowledge itself and its reflection in the human mind.
4. $ModelRating_4 > ModelRating_3 > ModelRating_2$?
   - In the 2^{nd} session, for 3 (out of 14) cases *VESA*^{1} provided the same solution as its human counterpart. For 24 out of 49 rating requests *VESA*^{1} provided the same rating as its human counterpart: $ident_2 = (3 + 24) = 27$, $ModelRating_2 := 27/53$. In the 3^{rd} session, for 17 (out of 28) cases *VESA*^{2} provided the same solution as its human counterpart. For 61 (out of 98) rating requests *VESA*^{2} provided the same rating as its human counterpart: $ident_3 = (17 + 61) = 79$, $ModelRating_3 := 79/126$. In the 4^{th} session, for only 8 (out of 28) cases *VESA*^{3} provided the same solution as its human counterpart. For 82 (out of 122) rating requests *VESA*^{3} provided the same rating as its human counterpart: $ident_4 = (8 + 82) = 90$, $ModelRating_4 := 90/150$.
   - With $ModelRating_4 = 0.6$, $ModelRating_3 \approx 0.63$, and $WellModeled_2 \approx 0.51$ we can at least claim that $ModelRating_4 \geq ModelRating_3 \geq ModelRating_2$ is almost met.
   - However, in the design of the experiment, a *VESA* was always based on former considerations of a present case by the same expert. A view on the decisions of the "most similar expert" showed, that this situation was better, when we had a setting where a former solution or rating is not available.
   - That these numbers are not convincing is due to the human factor in the experiment and the approach itself: All experts changed their opinion during the experiments for a remarkable number of cases. We believe the basic reasons are the interpretation of the cases itself and the fact that a solution often does not depend exclusively on the provided input attributes.
     In particular, the rating process of a *VESA* on the basis of a last consideration of this case in a solving (not rating) session is based on the assumption the domain is deterministic by nature, which is certainly not true for many interesting problem domains. This issue is discussed below.

Derived Improvements to VKB

Outdating Knowledge Since the number of solutions likely to be introduced in the rating process increases with the number of sessions, the probability to acquire some external knowledge increases over time. However, domain knowledge might become outdated. A strong indication for this fact is that a solution of *VKB* always receives bad marks. According to the basic philosophy that the recent human expertise is the primary and most reliable source of knowledge, an approach to face this problem is to remove entries that received bad marks for a long period.

Completion of VKB towards other than (former) test cases The fact that a *VKB* can only provide external knowledge (solutions) to cases that have been test cases in former validation sessions turned out to be a limitation of the practical value of the concept. The test cases for a current session are computed by analyzing the rules. They reflect the input–output behavior of the rule–based system and do not have, a priori, a big intersection with test cases of prior validation sessions that are in the *VKB*.

Derived Improvements to VESA

Consideration of alternative solutions Initially, we designed the *VESA*s in a way that they always consider both results of former solution sessions and results of former rating sessions. If, for example, a *VESA* is requested to rate a test case solution, and the currently missing expert considered this test input last in a solution session, *VESA* rated the lastly provided solution as "correct" and any alternative solution as "wrong". This is the assumption of a problem domain with unique solutions to each problem, which is not true for most interesting application fields of intelligent systems. Even if an expert prefers a particular solution when asked to solve a case, he/she might feel that alternative solutions are also fine and correct. This is in particular the case for problem classes like planning, scheduling, configuration, but also for some classification tasks. Thus, modeling a rating behavior by considering a previous solution behavior is not the right way. Vice–versa, the experiment also revealed that modeling a solving behavior based on a previous rating behavior is not appropriate. It seems to be quite arbitrary, which one of possibly several correct solutions has been rated most recently and thus, provided as *VESA*'s solution in the solving session. The experiment revealed that it is better to model an expert's solving behavior based only on former results of solving sessions and an expert's rating behavior based only on former results of rating sessions. Consequently, we refined the *VESA* concept to the one described here.

Computation of a most similar expert It turned out to be likely that the computation of a most similar expert results in several experts with the same degree of similarity with respect to their previous responses. In this case, we suggest to use the expert with the most recent identical behavior. This

seems to be reasonable, because the similarities in the behavior of humans are subject to natural change as well. This natural change can take place by different degrees and/or abilities to learn new insights.

Permanent validation of VESA The authors analyzed the experimentation results to validate *VESA*'s "validation knowledge". In fact, this validation needs to be performed by employing the *VESA*s all the time, even if its human source is available. By submitting *VESA*'s solution to the rating process of its human counterpart and comparing *VESA*'s rating with the one of its human counterpart, a *VESA* can easily be validated and statements about its quality can be derived: (1) The number of *VESA*'s solutions, which are rated by its human counterpart as "correct" (related to the total number of *VESA*'s solutions) and (2) the number of *VESA*'s ratings, which are identical with those of its human counterpart (related to the total number of *VESA*'s ratings) are session–associated validity degrees of a *VESA*'s solution– respectively rating ability.

Completion of VESA towards other than (former) test cases The fact that a *VESA* can only provide validation knowledge (solutions, ratings) to cases that have been test cases in former sessions turned out to be a limitation of the practical value of the concept. Test cases of an actual session are often different from test cases that have been considered in prior sessions. Following the intention of modeling the individual human expertise of its human source, the *VESA* approach needs to be refined by a concept of a "most likely" response of this human source in case there is no "most similar" expert who ever considered an actual case in the past. The authors' discussion of this issue did not reveal an approach that is mature to be published yet.

Summary

Application fields of intelligent systems are often characterized by having no other source of domain knowledge than human expertise. This source of knowledge, however, is often uncertain, undependable, contradictory, unstable, it changes over time, and furthermore, it is quite expensive. To address this problem, a validation framework has been developed that utilizes the "collective expertise" of an expert panel (Knauf et al. 2002).

However, even this approach does not yet utilize all opportunities to acquire human knowledge. With the objective of also using "historical knowledge" of previous validation sessions, a Validation Knowledge Base (*VKB*) has been introduced as a model of the "collective experience" of expert panels. Primary benefits are more reliable validation results by incorporating external knowledge and and/or a reduced need for current human input, for example smaller expert panels to reach the same quality of validation results. Furthermore, Validation Expert Software Agents (*VESA*) are introduced as a model of a particular expert's knowledge. Whereas the *VKB* can be considered (centralized) collective human expertise, a *VESA* can be considered a (decentralized) autonomous expertise, which is likely to be similar to the expertise of the modeled human counterpart. The *VKB* is more reliable, but may miss minor, yet possibly excellent human expertise. A *VESA*, on the other hand, can maintain such minor but possibly excellent human expertise.

A TURING Test experiment with a small prototype system indicates the usefulness of these concepts to model the collective (*VKB*) and individual (*VESA*) validation expertise. Generally, the idea of *VKB* is certainly the appropriate way to establish new sources of knowledge for system validation towards more reliable systems.

For both the *VKB* and the *VESA* concept, the experiments revealed some weaknesses of the approach derived refinements respectively research issues.

In fact, the experiment itself was a valuable source of knowledge. We gained many insights about the effects of our conceptual ideas and developed first refinement ideas towards AI systems with a better performance. The authors are convinced that the general approach of permanently checking the systems against cases derived from (historical and present) practice, is a necessary contribution to face the current problems of system dependability.

Acknowledgement

The first author would like to thank the Tokyo Denki University for inviting him to perform the experimental work behind this paper.

References

Knauf, R.; Gonzalez, A.J.; Abel, T. 2002. A Framework for Validation of Rule–Based Systems. *IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics*, vol. 32, # 3, pp. 281–295, 2004.

Knauf, R.; Tsuruta,S.; Ihara,H.; Gonzalez, A.J.; Kurbad,T. 2004a. Improving AI Systems' Dependability by Utilizing Historical Knowledge. *Proc. of 10^{th} International Symposium Pacific Rim Dependable Computing*, Papeete, Tahiti, French Polynesia, pp. 343-352, IEEE Computer Society Press, March 2004.

Knauf, R.; Tsuruta,S.; Ihara,H.; Kurbad,T. 2004b. Validating a Validation Technology: Towards a Prototype Validation Experiment. Invited talk. 45^{th} meeting of WG 10.4 of the International Federation for Information Processing (IFIP), Moorea Island, French Polynesia, March 2004.

Knauf, R.; Tsuruta,S.; Uehara,K.; Onoyama, T; Kurbad,T. 2004c. The Power of Experience: On the Usefulness of Validation Knowledge. *Proc. of 17^{th} International Florida Artificial Intelligence Research Society Conference 2004* (FLAIRS-2004), Miami Beach, FL, USA, pp. 337-342, ISBN 1-57735-201-7, Menlo Park, CA: AAAI Press, 2004.

Kurbad, T. 2003. A Concept to Apply a Turing Test Technology for System Validation that Utilizes External Validation Knowledge. Diploma Thesis, Ilmenau Technical University, Dept. of Computer Science and Automation, 2004.

Tsuruta, S.; Onoyama, T.; Taniguchi, Y. 2002. Knowledge–Based Validation Method for Validating Intelligent Systems. Kohlen (ed.): *Proc. of the 15^{th} Internat. Florida Artificial Intelligence Research Society Conference 2002 (FLAIRS-02)*, Pensacola Beach, FL, USA, pp. 226–230, Menlo Park, CA:AAAI Press.