PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1064 — 2005
Pozyskiwanie wiedzy i zarządzanie wiedzą

Rainer Knauf
Technical University Ilmenau, Germany
Setsuo Tsuruta
Tokyo Denki University, Japan
Avelino J. Gonzalez
University of Central Florida, USA

TOWARDS REDUCING HUMAN INVOLVEMENT IN VALIDATION OF KNOWLEDGE-BASED SYSTEMS

1. Introduction

In spite of significant advances, validation of knowledge-based systems (henceforth called *the system*) still requires the involvement of human experts. In contrast to verification, validation typically involves rigorous and extensive testing of the system. The results of these tests are nearly always evaluated by experts who may not always agree among themselves. The size of the test case set, the frequency of the validation exercises and the number of experts required for each such exercise can combine to pose great burdens of time and effort on experts. These experts are a scarce resource, have limited time, and are expensive to employ. These limitations have the potential to degrade a validation exercise.

Here, we describe an approach to managing the involvement of experts in the validation process. We address this with two different objectives in mind: 1) to improve the quality of the validation exercise, and 2) to reduce the dependence on human expert validators. We accomplish the former by maintaining persistent collective validation knowledge from previous validation exercises and the latter by introducing software agents that contain the validation knowledge of individual

expert validators. Our work is implemented in the context of our previously described *Validation Framework* [Knauf et al. 2002].

Briefly, the methodology incorporated in the Validation Framework provides a format for expert validator review of test cases and results using a variation of Buchanan and Shortliffe's Turing Test [1985] for system validation. In this step, humans play the role of experts validators as part of a *validation panel.* Their task is to 1) solve the test cases posed to the system under evaluation, 2) review and provide their judgment on the correctness (the *ratings*) of all anonymous solutions (the system's as well as the panel's own). Lastly, the Framework also provides an algebra to compute a *validity statement* as well as a means to refine its rules. This Framework is designed to work with rule-based systems, so we limit our discussion to such systems. Knauf [2000] provides a complete and detailed description of this Validation Framework.

Our contributions in this article 1) *a Validation Knowledge Base* (VKB) that contains collective validation knowledge from prior validation exercises, and 2) *Validation Expert Software Agents* (VESA) that can replace missing validators during validation exercises. The VKB can also help in reducing expert validator workload. Our concepts of a Validation Knowledge Base (VKB) and the Validation Expert Software Agents (VESA) were originally proposed by Tsuruta et al. [2000]. The work described here represents an extension of Tsuruta's work and its implementation in the context of the aforementioned Validation Framework. Tsuruta's work appears to be the first to specifically address the use of prior validation knowledge for improving the validation process. Their work aimed at developing validation solutions for commercial applications, and they address this issue frequently [Tsuruta et al. 2000].

Sections 0 and 0 describe these two concepts in earnest. Section 0 describes tests performed to evaluate their effectiveness when applied to a simple but non-trivial system. Section 0 discusses the results of the tests. In Section 0, we summarize the results and make conclusions about this work.

2. Collective Validation Knowledge – The Validation Knowledge Base

To improve the validation process, the *validation knowledge* used in prior exercises, namely the set of test cases (the test inputs and the best rated solutions) along with their authors, must persist from one validation exercise to the next. This is effectively accomplished by the Validation Knowledge Base (VKB). The VKB can also significantly reduce the involvement of expert validators by eliminating their need to *solve* old test cases whose solutions are already found in the VKB. The expert validator panel need only *solve* new test cases created by the Validation Framework that are not already part of the VKB. However, they still must *rate* all solutions. We continue by describing the internal structure of the VKB.

The VKB's (historical) test cases and their best rated solutions are described by 8-tuples [t_j, E_K, E_I, sol_{Kj}^{opt}, r_{IjK}, c_{IjK}, τ_S, D_C] where t_j is a test case input, sol_{Kj}^{opt} is a solution associated to t_j, which gained the maximum experts' approval, E_K is the list of experts who provided this solution, E_I is a list of experts who rated this solution, r_{IjK} is the rating of this solution, c_{IjK} is the certainty of this rating, τ_S is a time stamp indicating the validation session, and D_C is an informal description of any aspects of the test case that could not be described formally by the other seven elements. Additionally, a **list of supporters** $E_S \subseteq E_I$ for each solution sol_{Kj}^{opt} is kept in VKB. A supporter is rating expert, who provided a positive rating for sol_{Kj}^{opt} Table 1 shows how the VKB would appear for a simple application.

Table 1. Example Entries in VKB

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_3]	[e_1, e_2, e_3]	o_4	[1, 0, 1]	[1, 1, 1]	3	
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	
t_2	[e_3]	[e_1, e_2, e_3]	o_2	[1, 0, 1]	[1, 1, 1]	3	
t_2	[ ]	[e_1, e_2, e_3]	o_2	[1, 0, 1]	[1, 1, 1]	4	
t_3	[e_2]	[e_1, e_2, e_3]	o_{20}	[0, 1, 0]	[0, 1, 1]	1	
... ...							
t_{42}	[e_1, e_2, e_3]	[e_1, e_2, e_3]	o_{23}	[1, 1, 1]	[1, 1, 1]	2	
t_{42}	[e_1, e_2, e_3]	[e_1, e_2, e_3]	o_{23}	[1, 1, 1]	[1, 1, 1]	3	

Here, e_1, e_2 and e_3 are specific human expert validators, the outputs o_1, o_2,... are solutions, and the time stamps are denoted by natural numbers to indicate unspecified time when the validation exercise was held. The VKB is initially built as part of the first validation exercise. It is updated in subsequent validation exercises by adding all examined test cases.

Fig. 1 illustrates how the VKB fits into the Validation Framework. Here, the *test case generation* consists of two sub-steps (a) generating a quasi-exhaustive set of test inputs (*QuEST*) from the rule base structure and (b) reducing it down to a reasonably sized set of test cases (*ReST*) 0. Between these two sub-steps is the „entry-point" of the external validation knowledge in VKB. The *QuEST* for the current validation exercise and those cases in VKB that are duplicated by cases in *QuEST*, are both subjected to the reduction procedure that aims to build a subset of the test cases from both sources, *QuEST* and VKB. The procedure to form the reasonable set of test cases *ReST* now reduces the test cases contributed by the normally-generated *QuEST* in addition to the historical cases in the VKB that intersect the *QuEST*. The cases in VKB need to be included in the reduction process to 1) ensure that they meet the requirements of the current application and, 2) their ultimate number is small enough to avoid a time consuming and expensive test case experimentation.

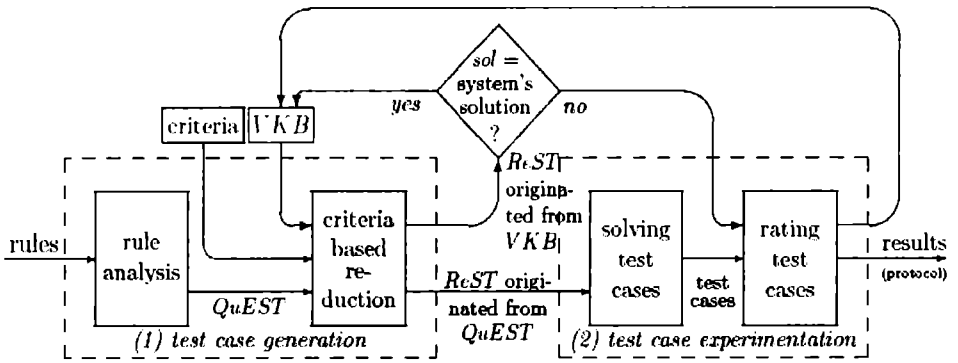

Fig. 1. The use of the VKB in the Validation Framework

The VKB can also help reduce the burden on the expert validators for the current exercise. The cases in the *ReST* that originate from the VKB already have their (historical best) solution in the VKB. Therefore, they don't have to be solved by the expert validators in the current process. Furthermore, not all of them may need to be rated, because the system's solution being in the rating process anyway and thus, a VKB's solution that is identical to it can be ignored.

The set of solutions $ExtSol \subseteq \Pi_2(ReST)$, which are contributed by the VKB and are subject to the rating process, is[1] $ExtSol := \{ sol: \exists Entry: Entry \in \mathbf{VKB}, \Pi_1(Entry) \in \Pi_1(\mathbf{ReST}), sol = \Pi_2(Entry) \}$. Because the criteria-based reduction process is controlled by a predetermined number m of cases that form the *ReST*, the *workload reduction factor* for the test case solving process for the expert validators can be quantified by the cardinality of $ExtSol$ (denoted by $|ExtSol|$) divided by $|ReST|$: **workload reduction factor** = $|ExtSol| / |ReST|$. The best-rated solutions associated with the test cases in the VKB represent an additional (external) source of expertise. The *expertise gain factor* introduced by the VKB is **expertise gain factor** = $|ReST| / (|ReST| - |ExtSol|)$.

3. Individual Validation Knowledge – The Concept of VESA

A Validation Expert Software Agent (VESA) is an intelligent agent corresponding to a particular expert validator. VESAs systematically model expert validators by keeping their validation knowledge and analyzing similarities with other expert validators. At some point, a VESA is able to serve as a temporary substitute for a human expert. The VKB is based on widely accepted knowledge and thus more reliable, but may miss obscure, yet possibly excellent human

[1] A bottom index i to a tuple denotes the i-th projection, i.e. the element at its i-th position. A bottom index i to a set of tuples denotes the set of the elements at the i-th position in each of the sets tuples.

expertise. A VESA, on the other hand, can model individual knowledge that is different from the accepted knowledge of the majority of expert validators. Thus, the VESA has the potential to maintain excellent and innovative individual human expertise.

It is assumed that e_i has still the same opinion about t_j's solution. If an expert validator e_i is not available to solve a case t_j, e_i's latest solution to t_j is provided by his/her VESA. If e_i never considered t_j before, similarities with other expert validators who might have the same "thinking structures" are considered. Among all expert validators who provided a solution to t_j, the one with the largest number of solutions equivalent to e_i's for the other cases is identified. e_i's solution is assumed to be the same as this other expert validator's and adopted by the VESA. Formally, a $VESA_i$ acts as follows when asked to provide a solution for a test case input t_j on behalf its expert validator e_i:

1. If e_i solved t_j in a prior validation exercise (with a value other than *unknown*), his/her solution with the latest time stamp τ_S will be furnished by $VESA_i$.

2. Otherwise,

a. All validators e_i, who ever delivered a solution to t_j form a set $Solver_i^0$, which is an initial dynamic agent for e_i: $Solver_i^0 := \{e' : [t_j, E_K, \ldots] \in VKB , e' \in E_K \}$

b. Select the most similar expert e_{sim} with the largest set of cases that have been solved equally by e_i and e_{sim} in the same prior validation exercise. e_{sim} forms a refined dynamic agent $Solver_i^1$ for e_i: $Solver_i^1 := e_{sim}: e_{sim} \in Solver_i^0 , \mid \{ [t_j, E_K, \_ , sol_{Kj}^{opt} , \_ , \_ , \tau_S, \_] : e_i \in E_K , e_{sim} \in E_K \} \mid \rightarrow max!$

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, $VESA_i$ provides *sol := unknown*.

If a $VESA_i$ is requested to provide a rating to a solution of a test case input t_j on behalf of expert validator e_i, it models the rating behavior of e_i as follows:

1. In case e_i rated t_j in a former session, $VESA_i$ adopts the rating with the latest time stamp τ_S and provide the same rating r and the same certainty c.

2. Otherwise,

a. All validators e_i, who ever delivered a rating to t_j form a set $Rater_i^0$, which is an initial dynamic agent for e_i: $Rater_i^0 := \{e' : [t_j, \ldots , E_I, \ldots] \in VKB , e' \in E_I \}$

b. Select the most similar expert e_{sim} with the largest set of cases that have been similarly rated by e_i and e_{sim} in the same session. e_{sim} forms a refined dynamic agent $Rater_i^1$ for e_i: $Rater_i^1 := e_{sim}: e_{sim} \in Rater_i^0 , \mid \{ [t_j, \_, E_I, sol_{Kj}^{opt} , r_{IjK} , \_, \tau_S, \_] e_i \in E_I , e_{sim} \in E_I \} \mid \rightarrow max!$

c. $VESA_i$ provides the latest rating r along with its certainty c to the present test case input t_j of e_{sim}.

3. If there is no such most similar expert e_{sim}, $VESA_i$ provides $r :=$ *norating* with a certainty $c := 0$.

Table 2 shows an example of a VESA 's behavior in a solving session that took place within our prototype experiment. The experiment compares a VESA's behavior ($VESA_2$, in the table) with the behavior of its human counterpart (e_2, in the table) to validate the VESA approach. t_i are test case inputs and o_i are the outputs provided by the VESA and its respective human expert validator.

Table 2. Examples for a VESA's solutions

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
t_{32}	o_8	o_3	t_{39}	o_9	o_9
t_{33}	o_8	o_8	t_{40}	o_{23}	o_{23}
t_{34}	o_2	o_2	t_{41}	o_{19}	o_{22}
t_{35}	o_8	o_8	t_{42}	o_{23}	o_{23}

EK_3 denotes the available "external knowledge" in the 3rd session, i.e. test cases with inputs of cases that have been solved and rated in the past. Here, in only one of the 14 test cases $VESA_2$ (the agent of the expert e_2) behaved different from its human counterpart. Table 3 shows an example of a VESA's behavior in a rating session that took place within the prototype experiment (see section 0). Again, EK_3 denotes the "external knowledge" within the 3rd session. Possible ratings are 1 (*correct solution*) and 0 (*incorrect solution*). Here, in 17 out of the 24 test cases $VESA_2$ (the representation of the expert validator e_2) behaved the same way as its human counterpart.

Table 3. Examples for a VESA's ratings

EK_3	Solution	Rating of		EK_3	Solution	Rating of	
		$VESA_2$	e_2			$VESA_2$	E_2
t_1	o_4	0	0	t_{29}	O_3	0	0
t_1	o_6	0	0	t_{29}	O_4	0	1
t_1	o_{21}	0	0	t_{29}	o_8	1	1
t_1	o_{18}	1	1	t_{29}	o_{16}	0	0
t_2	o_2	0	0	t_{30}	o_2	0	0
t_2	o_7	0	0	t_{30}	o_4	0	1
t_2	o_{20}	0	1	t_{30}	o_9	1	1
t_3	o_2	0	0	t_{30}	o_{16}	0	0
t_3	o_3	0	0	t_{31}	o_2	1	0
t_3	o_8	0	0	t_{31}	o_4	0	1
t_3	o_{20}	1	0	t_{31}	o_8	0	1
t_4	o_{23}	0	0	t_{31}	o_{16}	0	0

Actually, to learn the human experts' problem solving, VESA still depends on the knowledge of the expert validators. Learning involves analyzing the solving and rating performance of expert validators. The quality of learning and thus, the quality of VESA, depends on the quantity and coverage of data provided by the expert validators. VESA is able to replace its human source temporarily but it deteriorates if it does not acquire human input over a long period of time.

4. Evaluation of the VKB and VESA Concepts – A Test Prototype

To maximize the probability of enlisting adequate participation from expert validators for our experiment, we selected an amusing application: the selection of

an appropriate wine for a given dinner. We built a system by consulting the topical literature and deriving some informal knowledge from it.

The Knowledge Base: The problem 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]: $s_1 \in$ {*pork, beef, veal, venison,...*}, $s_2 \in$ {*raw, steamed, boiled,...*}, $s_3 \in$ {*Asian, Western*}}. The output O={o_1, ..., o_{24}} contains 24 different kinds of wine 0[2]: o_1= *Red wine, fruity, low tannin, less compound,* o_2= *Red wine, young, rich of tannin,* o_3= *Red wine, dark, fruity, from the new world* ... Expressing the informal knowledge as Horn clauses leads to a rule base R with 45 rules 0: 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 part of the Validation Framework 0 0, we computed *QuEST* from the rule base structure. It contained 145 cases 0 To reduce this to *ReST*, we applied domain related criteria and received 42 test inputs as the reasonable set of test cases *ReST* {t_1, ...,t_{42}}.

Application Conditions: Three human experts e_1, e_2 and e_3 and where available. The objective of our test program was to evaluate the feasibility of the VKB and VESA concepts, i.e. a) whether VKB can provide external knowledge from prior validation exercises to improve the validation process and b) whether VESAs can provide the same responses as their corresponding humans. In effect, we sought answers to the following questions:

1. Does the VKB increasingly contribute to the validation exercises in direct relation to the number of validation exercises? (*How many external solutions (outside the expertise of the current validation panel) are introduced into the rating process by the VKB?*)

2. Does the VKB increasingly contribute valid knowledge (best rated solutions) in direct relation to the number of validation exercises? (*How many of the VKB introduced solutions win the rating contest against the solutions of the current expert panel*?)

3. Does the VKB decreasingly acquire human expertise in direct relation to the number of validation exercises? (*How many new best-rated solutions are introduced into the VKB after a validation exercise? Does this decrease over time and several exercises*?)

4. Do VESAs' representations of their corresponding expert validator improve in direct relation to the number of validation exercises? (*Do VESAs provide the same solutions and ratings as their human counterparts?*)

We initially assumed empty VKB and VESAs. Therefore, each of the three experts and the wine selection system were asked to solve the 42 test cases in four separate sessions, each session representing a distinct validation exercise, possibly

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

separated in real life by several months or years. Each session consisted of 28 test cases, i.e. some test cases were solved more than once, as one would expect in an actual series of validation exercises, and were executed in the space of several days.[3] In session 2-4, one of the expert validators is modeled by its personal VESA. The session plan of the four sessions s shown in Table 4.

Table 4. Scheduled validation sessions

Session number	Experts			VESAs			Examined test case inputs out of $\Pi_I(ReST)$
	e_1	e_2	e_3	$VESA_1$	$VESA_2$	$VESA_3$	
1	+	+	+	-	-	-	$\Pi_I(ReST^1) := \{t_1, \ldots, t_{28}\}$
2	⊕	+	+	+	-	-	$\Pi_I(ReST^2) := \{t_{15}, \ldots, t_{42}\}$
3	+	⊕	+	-	+	-	$\Pi_I(ReST^3) := \{t_1, \ldots, t_{14}, t_{29}, \ldots, t_{42}\}$
4	+	+	⊕			+	$\Pi_I(ReST^4) := \{t_i : t_i \bmod 3 \neq 0\} = \{t_1, t_2, t_4, t_5, t_7, \ldots\}$

legend:
+ takes part in the session
  does not take part in the session
⊕ takes part in the session only for being compared with its VESA

$VESA_1$, **$VESA_2$**, and **$VESA_3$** model the individual expert validators e_1, e_2, and e_3 respectively. $\Pi_I(ReST^1)$, $\Pi_I(ReST^2)$, $\Pi_I(ReST^3)$, and $\Pi_I(ReST^4)$ are the subsets of the test case inputs from the entire *ReST*, which are used in the 1st, 2nd, 3rd and the 4th session. The entry ⊕ denotes that an expert, who is modeled by his/her VESA, is asked for solutions and ratings nevertheless to have the opportunity to compare the real experts' reply with the ones of their models. Each session leads to a progressively-updated VKB as well as updated VESAs for each of the expert validators e_1, e_2, and e_3.

For the VKB in each session, every best-rated solution sol_j^{opt} to a test input t_j is stored along with (a) a list of experts who provided this solution, (b) a list of experts, who provided ratings to this solution (c) their ratings and certainties, and (d) a time stamp. For the VESAs used in a session (indicated by "+" in Table 4) their behavior (i.e. their provided solutions and ratings) is computed as described in section 0.

We refer to the resulting VKBs and VESAs of an i^{th} session as VKB^i, $VESA_1^i$, $VESA_2^i$, and $VESA_3^i$. $ReST^i$ is the set of test cases computed for the current session i. Thus, the top index of *ReST* is larger than that of the VESAs by one.

For a fair evaluation of VKB, the intersection of the test case inputs found in VKB and those in *ReST* (***EK*** = external knowledge) needs to be considered in each validation session, along with those in the *ReST* but not in the VKB. EK^i denotes the external knowledge held by the VKB within the i-th experimentation session

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

(Table 4). Of course, EK^1 is empty, because there is no VKB built so far. All other EK^i are formed by the subset of $ReSi^i$ (the test case set of the i-th session), for which there are also solutions in VKB^i. The consideration of this external knowledge for the evaluation of the VESA concept is because this is the only knowledge that can be introduced into the rating process by the VKB from outside the current human expertise:[4] $EK^1 = \emptyset \cap \prod_1(ReST^1) = \emptyset$, $EK^2 = \prod_1(VKB^1) \cap \prod_1(ReST^2) = \{t_1, \quad t_{28}\} \cap \{t_{15}, \quad t_{42}\} = \{t_{15}, \quad , t_{28}\}$, $EK^3 = \prod_1(VKB^2) \cap \prod_1(ReST^3) = \{t_1, \ldots, t_{42}\} \cap \{t_1, \ldots, t_{14}, t_{29}, \ldots t_{42}\} = \{t_1, \ldots, t_{14}, t_{29}, \ldots t_{42}\}$, and $EK^4 = \prod_1(VKB^3) \cap \prod_1(ReST^4) = \{t_1, \ldots, t_{42}\} \cap \{t_1, t_2, t_4, t_5, t_7, t_8, \ldots\} = \{t_1, t_2, t_4, t_5, t_7, t_8, \ldots\}$. The cardinalities of these sets are $|EK^1| = 0$, $|EK^2| = 14$, $|EK^3| = |EK^4| = 28$.

We designed a set of metrics to address the four questions. After each session (session #i), beginning with the second session, we determine:[5]

- the number a_i of cases from VKB^{i-1} that were the subject of the rating session and relate it to $|EK^i|$ such that $A_1 := a_i / EK^i$
- the number b_i of cases from VKB^{i-1}, which provided the optimal (best rated) solution and relate it to $|EK^i|$ such that $B_1 := b_i / EK^i$
- the number c_i of cases from VKB^{i-1}, for which a new solution has been introduced into VKB and relate it to $|EK^i|$ such that $C_i := c_i / EK^i$
- the number d_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: $D_i := d_i$ ***/ required responses.***

The above four questions can now be re-addressed as follows in the context of these metrics: (1) $A_4 > A_3 > A_2$ *?*, (2) $B_4 > B_3 > B_2$ *?*, (3) $C_4 < C_3 < C_2$ *?*, and (4) $D_4 > D_3 > D_2$ *?*

5. Test Results and Discussion

Does the VKB increasingly contribute to the validation exercises in direct relation to the number of validation exercises ($A_4 > A_3 > A_2$)? With $A_4 \approx 0.85$, $A_3 \approx 0.071$, and $A_2 \approx 0.071$ the requirement $A_4 > A_3 > A_2$ was met at least in the step from the 3rd to the 4th session. The contribution effect could not really be expected as a result of the sessions before. The VKB needs to gain a certain amount of "historical experience" before it can contribute to a new session in a sufficient way. Indeed, after the third session, a remarkable number (24 out of 28 possible cases) of VKB^3 were introduced into the rating process.

Does the VKB increasingly contribute valid knowledge (best rated solutions) in direct relation to the number of validation exercises ($B_4 > B_3 > B_2$)? With $B_4 \approx 0.071$, $B_3 = 0$, and $B_2 = 0$, the requirement $B_4 > B_3 > B_2$ was also satisfied when

---

[4] $\prod_1(VKB^1)$ 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.

going from the 3rd to the 4th session. It is remarkable that in the 4th session, VKB[3] contributed solutions for two cases that had not been provided by the human experts, but won the rating "contest". This is exactly the intended effect of the VKB -introducing new knowledge that turned out to be more valid than the knowledge provided by the current panel. In fact, the more entries a VKB gains, the higher the number of solutions that are the subject of the rating process. Thus, with an increasing number of sessions, the probability that a VKB contributes such (more valid) knowledge than the human expert validators increases.

Does the VKB decreasingly acquire human expertise in direct relation to the number of validation exercises ($C_4 < C_3 < C_2$)? With $C_4 \approx 0.61$, $C_3 \approx 0.57$, and $C_2 = 0.5$ the requirement $C_4 < C_3 < C_2$ was not met. We probably asked the wrong question. The underlying assumption for this question is a static problem domain with a static domain knowledge that needs to be explored systematically. This was not true for the considered domain. We believe that in interesting problem domains, the domain knowledge itself as well as how it is interpreted by humans changes over time.

Do VESAs' representation of their corresponding expert validator improve in direct relation to the number of validation exercises ($D_4 > D_3 > D_2$)? With $D_4 = 0.6$, $D_3 \approx 0.62$, and $D_2 \approx 0.43$ at least $D_4 \geq D_3 \geq D_2$ is nearly met. In 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" in [Knauf et al. 2004] shows that this situation was better when we had a situation where a former solution or rating is not available.

Generally, the tests indicated that the VKB and, to a lesser degree, the VESA concepts are indeed feasible and useful in the validation process. However, our results also gave us some reason to pause, especially as it relates to VESA. Here are the lessons learnt:

Experimentation Planning: In the experiment, the VKB increased the number of solutions to be rated significantly with increasing number of sessions. This is caused by the conditions of the experiments, in which the System Refinement step of the Validation Framework[6] was omitted. This caused the system solution to never improve with the results of each validation cycle. A solution that turned out to be invalid was never replaced in the rule base and thus, a subject of validation in each session.

Outdating Knowledge: Domain knowledge might become outdated. A strong indication of this would be when a solution contributed by the VKB repeatedly receives poor ratings whenever it is introduced in a rating session. One approach to this problem is to analyze the prior ratings of each entry in the VKB and remove those entries that have received poor ratings for an arbitrary extended period.

---

[6] This was because of (a) time limitation for the process of the experiment and (b) non-relevance to the purpose of the experiment.

Completion of VKB towards other than (former) test cases: The fact that a VKB can only provide external knowledge to cases that have been test cases in former validation sessions turned out to be a limitation on the practical value of the concept. The test cases for a current session are computed by the test case generation algorithms which uses the rule base structure. In situations where significant changes to the rules have been made, the newly developed *ReST* may have relatively few test cases in common with test cases found in the VKB. For this reason, the VKB has only limited applicability for validation exercises where significant changes to the system are being evaluated.

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. This did not happen in our experiment when determining the reply of a VESA to a request for a solution or a rating (see section 0). However, the computed similarities were similar enough that we can foresee this being a recurrent problem. In such cases, we suggest using the expert with the most recent identical behavior to maximize the probability that the latest thinking is employed.

Continuous validation of VESA: The authors analyzed the experiment results to validate VESA's validation knowledge. This continuous validation of the VESAs should be performed by employing a VESA in the background at all times when its human counterparts are available. By (a) submitting VESA's solution to the rating process of its human counterpart and (b) comparing VESA's rating with that of its human counterpart, a VESA can be validated and statements about its quality can be derived.

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

6. Summary and Conclusion

Application fields of knowledge-based systems are characterized by having no other source of domain knowledge than human expertise. This source of knowledge has several drawbacks: It is often uncertain, undependable, contradictory, and unstable; it changes over time, and is quite expensive. To address this problem, a Validation Framework 0, 0 utilizes the collective expertise

of an expert validator panel. This way, the human knowledge used for validation becomes impartial.

However, this approach does not utilize all opportunities to acquire and employ human knowledge in system validation. With the objective of also using historical knowledge of previous validation sessions, a Validation Knowledge Base (VKB) has been introduced in this article. It can be considered to be a representation of the collective experience of expert panels that participated in previous validation exercises. A VKB is constructed and maintained across various validation exercises. Primary benefits are (a) more reliable validation results by incorporating external knowledge and and/or (b) a reduced need for current human input, for example smaller expert panels to reach the same quality of validation results. Additionally, (c) it can be used to improve the selection of an appropriate expert panel based on their prior performances, and (d) improve the identification of an optimal solution to test cases.

Furthermore, Validation Expert Software Agents (VESA) one for each particular expert validator of a current validation session - are introduced. They are a representation of a particular expert's individual knowledge. A VESA systematically reproduces the validation knowledge and behavior of its human counterpart by retrieving the expert validator's previous solution and/or rating of a test case. If the expert validator did not address the same test case in prior exercises, then it analyzes similarities with the responses of other expert validators and retrieves the one from the most similar expert. After a learning period, it can be used to substitute the human expert temporarily.

Whereas the VKB can be considered (centralized) collective human expertise, a VESA can be considered (decentralized) individual expertise and is likely to be representative of the expertise of its human counterpart. The VKB is more reliable, but may miss obscure, yet possibly excellent human expertise. A VESA, on the other hand, can maintain such obscure but possibly excellent human expertise.

An experiment with a small prototype system indicates the usefulness of these concepts to model the collective (VKB) and individual (VESA) validation expertise. Generally, VKB proved to be an appropriate way to establish new sources of knowledge for system validation. The tests demonstrated its ability to provide external knowledge that can be useful in the validation exercise even when a panel of expert validators was available. The experiments revealed some weaknesses of the VESA approach. The experiment itself was a valuable source of knowledge. We gained insights about the effects of our conceptual ideas and developed refinement ideas towards AI systems with a better performance. We are convinced that the general approach of permanently checking the systems against cases derived from practice is a necessary contribution to face the current problems of system dependability.

References

Buchanan B.G., Shortliffe E.H., *Rule-Based Expert Systems – The NYCIN Experiments of the Stanford Heuristic Programming Project*, Reading, MA: Addison Wesley, 1985.

Knauf R., *Validating Rule-Based Systems – A Complete Methodology*, Aachen: Shaker, 2000 (Berichte aus der Informatik) Zugl. Ilmenau, Technische Universität, Habilitationsschrift, 2000.

Knauf R., Gonzalez A.J., Abel T., *A Framework for Validation of Rule-Based Systems*, IEEE Transactions of Systems, Man and Cybernetics - Part B: Cybernetics, Volume 32, No. 3, June 2002, pp. 281-295, 2002.

Knauf R., Tsuruta S., Uehara K., Gonzalez, A.J., *Validation Knowledge Bases and Validation Expert Software Agents. Models of Collective and Individual Human Expertise*, Technical Report, Tokyo Denki University, School of Information Environment, 94p., available at http://www.theoinf.tu-ilmenau.de/ki/ki/veroeff.html, Tokyo, Japan, 2004.

Tsuruta S., Onoyama T., Kubota S., Oyanagi K., *Validation Method for Intelligent Systems*, Proceedings of 13th International Florida Artificial Intelligence Research Society Conference, (FLAIRS-2000), pp. 361-365, Orlando, FL, 2000.

O REDUKOWANIU ZAANGAŻOWANIA CZYNNIKA LUDZKIEGO W WARTOŚCIOWANIE SYSTEMÓW Z BAZĄ WIEDZY

Streszczenie

Eksperci zaangażowani w zadania obejmujące wartościowanie wiedzy w systemach z bazą wiedzy (eksperci wartościujący) często mają ograniczone możliwości czasowe i niezadowalająca jest ich dostępność. Co więcej, mają oni często różne opinie czy zmieniają je z upływem czasu. Chcemy poprawić tę sytuację poprzez wartościowanie wiedzy stosując poprzednia wykonane zadania dotyczące wartościowania dla tego samego systemu. Prezentujemy Wartościującą Bazę Wiedzy (WBB), która zawiera zbiór poprzednio wykonanych zadań związanych z wartościowaniem przez najlepszych ekspertów. Podstawowa korzyść polega na bardziej wiarygodnych wynikach wartościowania i obniżenia pracochłonności ekspertów. Prezentujemy także koncepcję agentów wartościujących systemy ekspertowe (AWSE), którzy reprezentują szczegółową wiedzę ekspertów. Po pewnym okresie uczenia, baza jest systematycznie wymieniana przez wiedzę ekspertów. Pomaga to w redukowaniu zaangażowania ekspertów lub utrzymania ich oczekiwanej aktywności w sytuacji, kiedy nie są dostępni. Opisujemy także eksperymenty z małym prototypem systemu oceniającego te koncepcje.