Human Expertise in Fault Detection and Adjustment

An Empirical Case Study

Rainer Knauf

Technical University of Ilmenau
School of Computer Science and Automation
Ilmenau, Germany

Setsuo Tsuruta

Tokyo Denki University
School of Information Environment
Tokyo, Japan

Avelino J.Gonzalez

University of Central Florida
Dept. of Electrical and Computer Engineering
Orlando, FL, USA

Content

1. System Evaluation and Refinement – An Issue of this WG?
2. Our Concept – An Overview
3. The Problem with Human Experience

Incorporating a Validation Knowledge Base (VKB) as a Model of Collective Experience

Incorporating Validation Expert Software Agents (VESA) as Models of Individual Experiences

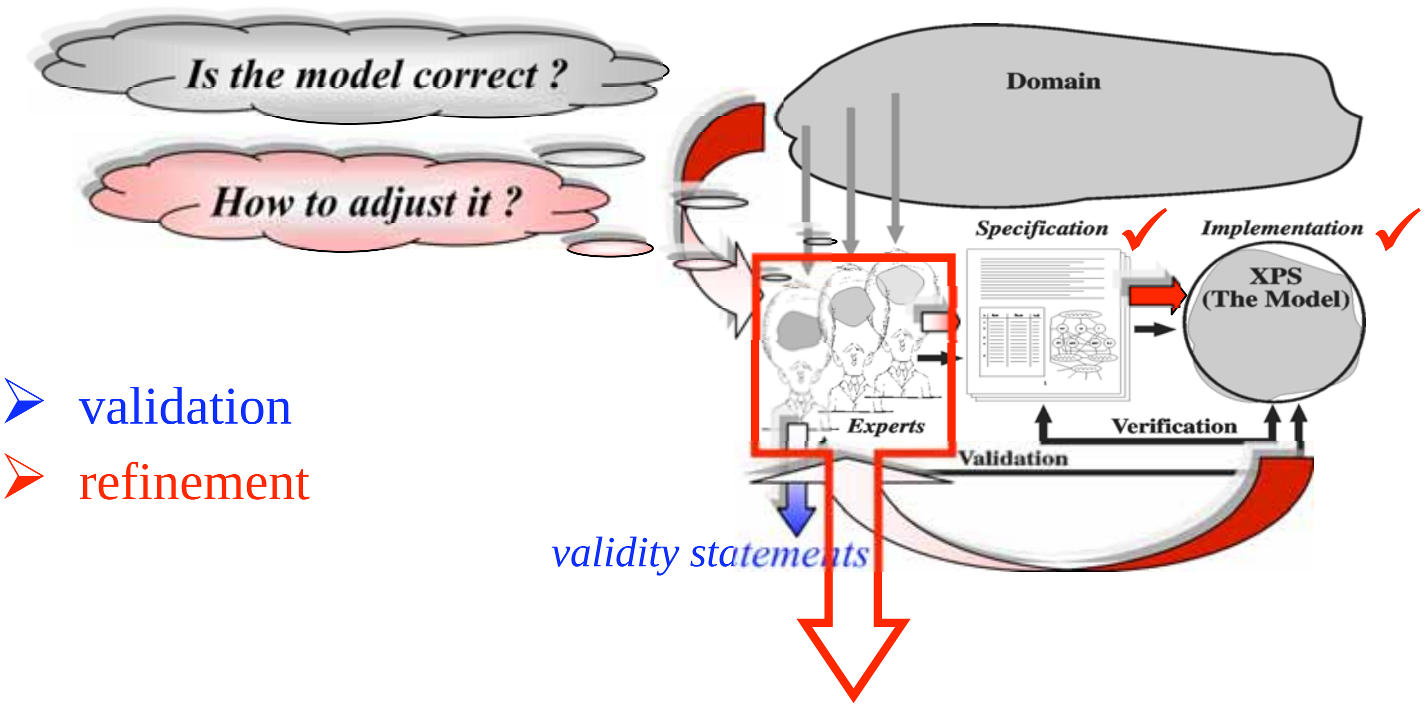
6. A Prototype Test
   - *Knowledge Base*
   - *Test Cases*
   - *Application Conditions*
7. Test Results
   - *On the Usefulness of Modeling the experience*
   - *Lessons Learnt*
8. Summary and Conclusion

1. System Evaluation and Refinement – An Issue of this WG?

- *Today's opportunities to design and employ complex systems rise the question, whether or not we are able to control what we are able to build*
- *The impact of invalidity increases with the with the number today‘s systems‘ application fields and their sensibility to malfunctions*
- *Today‘s IT-systems may become a real threat without ensuring their validity*
- *Moreover, many interesting applications are characterized by some dynamics in their topical background.*
- *Thus, these systems need to be refined based on both, revealed invalidities and new topical insights.*

- In fact, these concerns are issues of **dependable computing**.
- Maybe they are not an issue of fault tolerance, but of **fault detection and adjustment** instead.

Verification, validation, and refinement – what‘s it?



- validation
- refinement

Humans in the loop – a problem?

Yes, indeed!

But is there any alternative ?

2. Our Conceept – An Overview

Step # 1: **Test case generation**
Generate and optimize a set of test cases ***[test data , expected output]***
that meets the competing requirements (1) ***coverage*** *and* (2) ***efficiency***

Step # 2: **Test case experimentation**
Exercise the test data by both the system under investigation and a panel of validating experts as a TURING *Test - like experiment*

Step # 3: **Evaluation**
Interpret experimentation results & report test case associated invalidities

Step # 4: **Validity assessment**
Analyze reported results and conclude validity assessments associated with (1) *test cases,* (2) *outputs,* (3) *rules, and* (4) *the entire system*

Step # 5: **System refinement**
Formally reconstructing the rule base so that it infers best rated solutions

3 The Problem with Human Experience

What's the problem with employing human expertise for system validation?

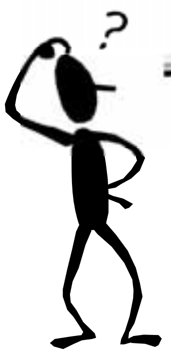
☹ Experts have different beliefs, experiences and learning capabilities.

☹ Experts are not free of mistakes.

☹ Experts' opinions about the desired system's behavior

- differ from each other
- change over time as a result of misinterpretations, mistakes or new insights

☹ Experts are often too busy and/or too expensive to hire them for system validation and refinement.



How to get out of this misery ?

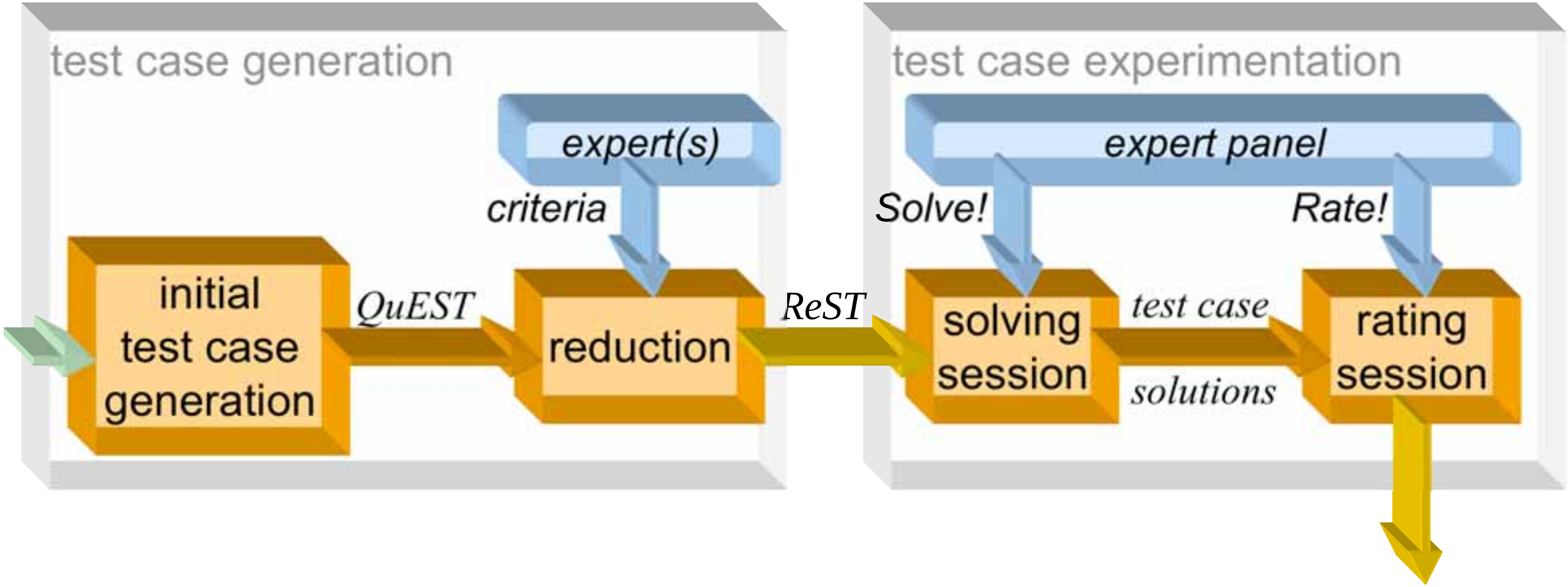
By

(1) **modeling their experience**

(2) **compensating some human weaknesses with this model**

The Involvement of Humans so far

Where is the human input into our validation technology ?



QuEST **Qu**asi **E**xhaustive **S**et of **T**est Cases

- ➢ *a well-designed set that ensures coverage by formally analyzing the input space*

ReST **Re**asonable **S**et of **T**est Cases

- ➢ *a subset of **QuEST** that ensures the requirement efficiency by using validation criteria*

Objectives of modeling human experience

Supplementing additional expertise to the validation panel, in particular:

- Suggesting new solutions to test cases, different from the panel's suggestions
- Offering additional input without consulting humans
- Substituting missing individual human expertise
- *… others $\notin$ this talk*

4 Incorporating a Validation Knowledge Base (VKB) as a Model of Collective Experience

4.1 The Content of VKB

All formal and informal data that can be collected, i.e. to each test case

- the (input) test data t_j
- a list of all solvers E_{Kj}
- a list of all raters E_{Ij}
- associated optimal (best rated) solution sol_{Kj}^{opt}
- the ratings provided by the rating experts r_{IjK}
- the certainties of these ratings c_{IjK}
- a session time stamp τ
- an informal description of the context D_j

Thus, ***VKB*** is a set of 8-tuples $[\, t_j\,, E_{Kj}, E_{Ij}, sol_{Kj}^{opt}, r_{IjK}, c_{IjK}, \tau, D_j \,]$

A part of *VKB* in the prototype test experiment

e_1, e_2, e_3
human experts

$t_1, t_2, \ldots$
test case inputs

$o_1, o_2, \ldots$
solutions (outputs)

τ
session #

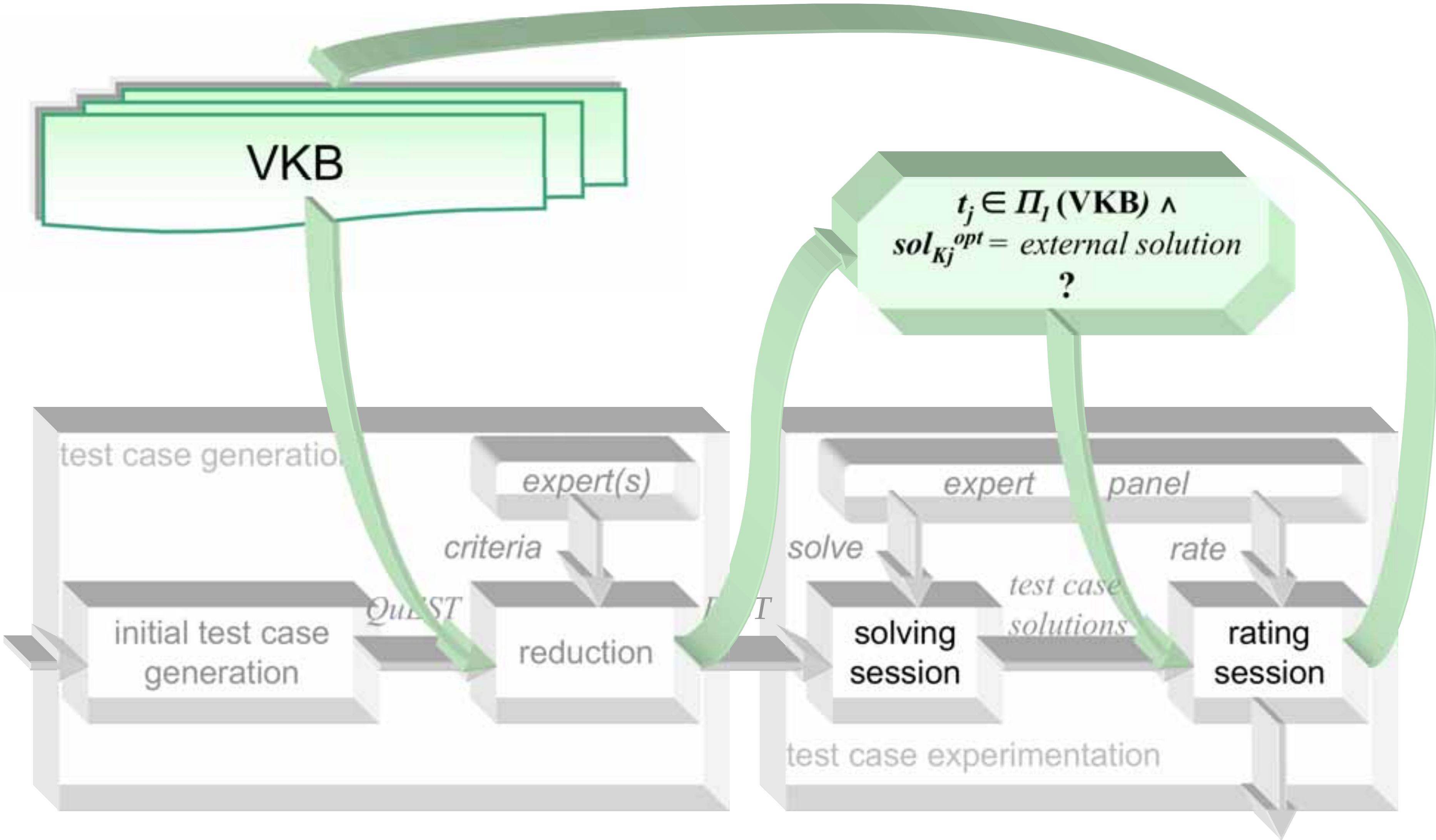
r
rating: 1 for correct, 0 for incorrect

c
certainty: 1 for certain, 0 for uncertain

t_j	E_{Kj}	E_{Ij}	sol_{Kj}^{opt}	r_{IjK}	c_{IjK}	τ	D_j
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	

4.2 The Usage of VKB

External collective experience: $sol \in$ VKB, but not provided by the panel



Quantifying the supplement of VKB to the human expertise

Set of external solutions (not provided by the current panel):

$$ExtSol := \{\, sol \colon \exists\, Entry \colon Entry \in VKB,\ \Pi_1(Entry) \in \Pi_1(\mathrm{ReST}),\ sol = \Pi_4(Entry) \,\}$$

⇒ **Workload reduction factor of the VKB**

- *by skipping the solving process*

$$\textit{workload reduction factor} = |\, ExtSol \,| \,/\, |\, ReST \,|$$

⇒ **Expertise gain factor of the VKB**

- *by supplementing ReST with interesting solutions outside the panel's expertise*

$$\textit{expertise gain factor} = |\, ReST \,| \,/\, (\, |\, ReST \,| - |\, ExtSol \,| \,)$$

5 Incorporating Validation Expert Software Agents (VESA) as Models of Individual Experiences

Objectives

- Forming a model of each validator‘s individual knowledge and behavior
- Successive refinement of this model by consecutive validation sessions

Source of *VESA*‘s knowledge: solving and rating results

- of the associated human counterpart
- of other human validators who often have the same opinion as the associated human origin

VESAs

- *are formed just in the moment of their need and „forgotten“ after their usage*
- *model just the required aspect of their human origin based on historical information of former sessions (i.e. not the current session)*
- *are requested in case its human counterpart is not available*
- *may be requested even if the human origin is present to validate the VESA concept itself by comparing the behavior of VESA with the real one of the human source.*

VESA models the solving behavior of an expert e_i for a test case t_j as follows

Step # 1

In case e_i solved (*with a solution different from „unknown“*) t_j in a former session, his/her solution with the latest time stamp τ will be provided by ***VESA***.

Step # 2

- ✓ 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 : $Solver_i^0 := \{e' : [t_j, E_{Kj}, ...] \in VKB \wedge e' \in E_{Kj}\}$
- ✓ 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 in the same session. e_{sim} forms a refined dynamic agent $Solver_i^1$ for e_i :

$$Solver_i^1 := e_{sim} : (e_{sim} \in Solver_i^0) \wedge (|\{[t_j, E_{Kj}, \_, sol_{Kj}^{opt}, \_, \_, \tau, \_] : e_i \in E_{Kj}, e_{sim} \in E_{Kj}\}| \rightarrow \max!)$$

- ✓ Provide the latest solution of the expert e_{sim} to t_j , i.e. the solution with the latest time stamp τ by ***VESA***.

Step # 3

If there is no such most similar expert, provide the solution $sol := unknown$ by ***VESA***.

An example of a *VESA* 's solving behavior compared to the human counterpart

EK^3

external knowledge (entries of the VKB) available in the 3rd session

e_2

human expert #2

$t_1, t_2, \ldots$

test case inputs

$o_1, o_2, \ldots$

solutions (outputs)

$VESA_2$

the VESA-model of expert #2

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}

VESA models the rating behavior of an expert e_i for a test case t_j as follows

Step # 1

In case e_i rated t_j in a former session, adopt the rating with the latest time stamp τ_s and provide the same rating r and the same certainty c by ***VESA***.

Step # 2

- ✓ 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 : $Rater_i^0 := \{e' : [t_j, \_, E_{Ij}, \ldots] \in VKB \wedge e' \in E_{Ij}\}$

- ✓ 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 in the same session. e_{sim} forms a refined dynamic agent $Rater_i^1$ for e_i :

$$Rater_i^1 := e_{sim} : (e_{sim} \in Rater_i^0) \wedge (|\{[t_j, \_, E_{Ij}, sol_{Kj}^{opt}, r_{IjK}, \_, \tau, \_] : e_i \in E_{Ij}, e_{sim} \in E_{Ij}\}| \rightarrow \max!)$$

- ✓ Provide the latest rating r of the expert e_{sim} along with its certainty c, i.e. the ones with the latest time stamp τ , to the present test case t_j by ***VESA***.

Step # 3

If there is no such most similar expert, provide the rating $r := norating$ along with a certainty $c := 0$ by ***VESA***.

An example of a *VESA* ‘s rating behavior compared to the human counterpart

EK^3
: *external knowledge (entries of the VKB) available in the 3rd session*

e_2
: *human expert #2*

$t_1, t_2, ...$
: *test case inputs*

$o_1, o_2, ...$
: *solutions (outputs)*

$VESA_2$
: the VESA-model of expert #2

EK_3	solution	rating of $VESA_2$	rating of e_2	EK_3	solution	rating of $VESA_2$	rating of 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

6 A Prototype Test

How to find human experts who are able and willing to cooperate for free ?

By choosing an "application" with a certain "entertainment factor":

Selection of an appropriate wine for a given dinner

6.1 The Knowledge Base

Input space: $I := [\, s_1, s_2, s_3 \,]$:

- $s_1 \in \{$ *pork, beef, veal, fowl,…, fish,…,goat cheese,…, fruit dessert, ice cream* $\}$
- $s_2 \in \{$ *non(raw), steamed, boiled, grillesd, fried,* … $\}$
- $s_3 \in \{$ *Asian, Western* $\}$

Output space: $O := \{\, o_1, o_2, \ldots, o_{24} \,\}$ with

- o_1 = *Red wine, fruity, low tannin, less compound*
- o_2 = *Red wine, young, rich of tannin*
- …

Rule base: $R := \{\, r_1, r_2, \ldots, r_{45} \,\}$ with

- $r_1 : o_1 \leftarrow (s_1 = \textit{fowl})$
- $r_2 : o_1 \leftarrow (s_1 = \textit{veal})$
- $r_3 : o_2 \leftarrow (s_1 = \textit{pork}) \wedge (s_2 = \textit{grilled})$
- …

6.2 The Test Cases

... have been generated with a technology as introduced in former papers.

The resulting “Reasonable Set of Test Cases” (***ReST***) is:

t_1	pork	boiled	Asian	t_{22}	fish	steamed	Western
t_2	pork	grilled	any	t_{23}	fish	boiled	Asian
t_3	pork	fried	any	t_{24}	fish	grilled	any
t_4	pork	stewed	any	t_{25}	fish	fried	any
t_5	beef	boiled	Asian	t_{26}	fish	stewed	Asian
t_6	beef	grilled	any	t_{27}	fish	deep fried	Asian
t_7	beef	fried	any	t_{28}	hard cheese	non	Western
t_8	beef	stewed	any	t_{29}	hard cheese	casserole	Western
t_9	veal	boiled	any	t_{30}	hard cheese	deep fried	Western
t_{10}	veal	grilled	any	t_{31}	soft cheese	non	Western
t_{11}	veal	fried	any	t_{32}	soft cheese	casserole	Western
t_{12}	veal	stewed	any	t_{33}	soft cheese	deep fried	Western
t_{13}	venison	boiled	any	t_{34}	goat cheese	non	Western
t_{14}	venison	grilled	any	t_{35}	goat cheese	casserole	Western
t_{15}	venison	fried	any	t_{36}	goat cheese	deep fried	Western
t_{16}	venison	stewed	any	t_{37}	blue mold cheese	non	Western
t_{17}	fowl	boiled	any	t_{38}	blue mold cheese	casserole	Western
t_{18}	fowl	grilled	any	t_{39}	blue mold cheese	deep fried	Western
t_{19}	fowl	fried	any	t_{40}	fruit dessert	non	any
t_{20}	fowl	stewed	any	t_{41}	aromatic dessert	non	any
t_{21}	fish	non	Asian	t_{42}	ice cream	non	any

6.3 Application Conditions

The experimentation took place with

- three human experts e_1 , e_2 , e_3
- a test case set ***ReST*** = { t_1 , t_2 , …, t_{42} }
- session schedule:

session number	experts			VESAs			examined test case inputs out of $\Pi_1(ReST)$
	e_1	e_2	e_3	$VESA_1$	$VESA_2$	$VESA_3$	
1	+	+	+	-	-	-	$\Pi_1(ReST^1) := \{ t_1, \dots, t_{28} \}$
2	⊕	+	+	+	-	-	$\Pi_1(ReST^2) := \{ t_{15}, \dots, t_{42} \}$
3	+	⊕	+	-	+	-	$\Pi_1(ReST^3) := \{ t_1, \dots, t_{14}, t_{29}, \dots, t_{42} \}$
4	+	+	⊕	-	-	+	$\Pi_1(ReST^4) := \{ t_i : t_i \bmod 3 \neq 0 \}$

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

<u>*Notational Conventions*</u>

- ***VKB^i*** *denotes the* ***VKB*** *as developed after the* ***i*** *-th session*
- ***$VESA_k^i$*** *denotes the behavior of the* ***VESA*** *which models the behavior of expert* ***e_k*** *after the* ***i*** *-th session*
- ***$ReST^i$*** *denotes the test case set used in the* ***i*** *-th session*
- ***EK^i*** *denotes the available "external knowledge" of the* ***VKB*** *in the* ***i*** *-th session:* ***$EK^i := \Pi_1(VKB^i) \cap ReST^i$***

6.4 Desired Outcome of the Experiment

The experiment should provide 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 models of their human source improve with in increasing number of validation sessions?*
   - ***Do the VESAs provide the same solutions and ratings as their human counterpart?***

To quantify these measures, we computed after each session (session # i)

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

Thus, desired answers can be formalized

1. *Does the VKB contribute to the validation sessions at an increasing rate with an increasing number of validation sessions:* $A_4 > A_3 > A_2$?
2. *Does the VKB contribute valid knowledge (best rated solutions) in an increasing rate with an increasing number of validation sessions:* $B_4 > B_3 > B_2$?
3. *Does the VKB increasingly gain the human expertise as number of validation sessions increases:* $C_2 > C_3 > C_4$?
4. *Do the VESAs model of their human source improve with in increasing number of validation sessions:* $D_4 > D_3 > D_2$?

7 Test Results

Does the VKB contribute to the validation sessions at an increasing rate with an increasing number of validation sessions: $A_4 > A_3 > A_2$?

- \# of new external solutions from VKB:
  - 1 (of 14 possible in EK) in session 2
  - 2 (of 28) in session 3
  - 24 (!) (of 28) in session 4 $0.85 >> 0.071 \geq 0.071$
- *Obviously, the VKB needs to gain some “initial experience” before it contributes a remarkable number of new solutions.*
- *The desired effect became remarkable in the 4th session.*

2. *Does the VKB contribute valid knowledge (best rated solutions) in an increasing rate with an increasing number of validation sessions:* $B_4 > B_3 > B_2$?

- \# of new external solutions, which won the rating session:
  - 0 (out of 14) in session 2
  - 0 (out of 28) in session 3
  - 2 (out of 28) in session 4: $0.071 \geq 0 \geq 0$
- *However, it is remarkable that 2 solutions which were not provided by the panel got very best marks by the same panel.*
- *This is what we want the VKB to do: Contributing better knowledge than the current human experts. The „collective experience“ of former panels reveals to be better than the current panel.*

3. *Does the VKB increasingly gain the human expertise as number of validation sessions increases:* $C_2 > C_3 > C_4$?

   - # of cases introduced into VKB:
     - 7 (of 14) after session 2
     - 16 (of 28) after session 3
     - 17 (of 28) after session 4: $0.5 \leq 0.57 \leq 0.61$
   - *Here, our expectation was not met!*
   - *The reason is probably, that the domain knowledge itself as well as its reflection in human minds changed from session to session.*
   - *Most interesting problem domains are not static by nature; individual peoples' opinions are not static by nature.*

4. *Do the* ***VESA****s model of their human source improve with in increasing number of validation sessions:* $D_4 > D_3 > D_2$?

   - # of identical responses by the expert and his/her VESA
     - 27 (of 63) in session 2
     - 78 (of 126) in session 3
     - 90 (of 150) in session 4: $0.6 \approx 0.62 > 0.43$
   - *Again, we explain this as the result of changing minds by the experts.*
   - *A crucial problem is*
     - *the interpretation of a verbal case description and*
     - *some latent dependence from other circumstances than the case input itself (the mood, e.g.).*

Lessons Learnt

Derived improvements to the „collective experience“ in ***VKB***

- ✓ Outdating knowledge
  - ➢ *Should some knowledge, which receives „bad marks“ by several expert panels over many sessions removed from VKB?*
- ✓ Completion of VKB towards other than former test cases
  - ➢ *VKB so far can only provide its „experience“ only for historic cases.*
  - ➢ *How to derive experience from VKB for other cases? Is a CBR concept appropriate for this problem?*
  - ➢ *Current work: Adapting the k-NN Data Mining Approach towards solving this problem*

Derived improvements to the „individual experience“ in ***VESA*s**

- ✓ Non-deterministic problem domains
  - ➢ *A certain solution might be „correct“ in the eyes of an expert, even if it is not the one he would provide as a solution to the presented case.*
  - ➢ *In many interesting problem domains cases have several acceptable solutions.*
  - ➢ *This drawback has already been fixed:*
    - *VESA's solving behavior is modeled based only on the solving behavior of its human counterpart.*
    - *VESA's rating behavior is modeled based only on the rating behavior of its human counterpart.*
- ✓ Determination of a „most similar expert“
  - ➢ *The prototype experiment revealed, that there are often several experts' solution in the VKB with the same degree of similarity.*
  - ➢ *In this case we suggest to consider another parameter: We should look for an expert with the <u>most recent</u> identical (solving or rating) behavior.*
  - ➢ *This is reasonable, because also such similarities are subject to natural change over time.*

Derived improvements to the „individual experience" in ***VESA*s** *(cont'd)*

- ✓ Permanent validation of the ***VESA*s**
  - ➢ *The concept will be refined by adding some permanent „self-validation" of each VESA by*
    - ➢ *submitting VESA's solution to the rating process of its human counterpart and*
    - ➢ *comparing VESA's rating with the rating of its human counterpart.*
  - ➢ *Thus, some statement about each VESA's quality can be derived:*
    - *The number of VESA's solutions, which are rated by its human counterpart as „correct" and*
    - *the number of VESA's ratings which are identical with those of its human counterpart*

    are measures about the performance of the human behavior model.
- ✓ Completion of ***VESA*s** towards other than former test cases
  - ➢ *In case there is no „most similar expert" who ever considered (solved or rated) a current case, a concept of determining a „most likely response" of the modeled expert needs to be developed.*

8 Summary and Conclusion

1. Ensuring validity of AI systems requests methods beyond conventional software engineering techniques. The only source of domain knowledge is often human expertise.

2. Human expertise is often uncertain, undependable, contradictory, unstable, it changes over time and is quite expensive.

3. The concept of ***VKB*** is the key to use this resource more efficiently towards valid systems. The VKB approach includes all aspects of „collective historical experience“ that have been provided by previous expert panels.

4. While ***VKB*** aims at modeling the human experts‘ collective and most accepted (best rated) knowledge, the ***VESA*** concept aims at modeling the individual human experts.

5. Experiments revealed that the ***VKB*** and ***VESA*** approach needs to be refined with respect to

   - their completion towards other than (previous) test cases

     Under construction: Adapting the k-NN data mining approach

   and ***VESA*** needed to be developed further with respect to

   - the nature of the non-deterministic problem domains (done!)

     Solving cases based on a previous rating is not appropriate

   - their permanent validation