Trust Put to the Test: a Testcase for a Cognitive Trust Model

Juri Luca De Coi, Laurent Vercouter

Lyon, France, 25-08-2011

Outline

3. Trust Put to the Test
2. a Testcase…
1. … for a Cognitive Trust Model
4. Implementation
5. Conclusions & further work

Outline

3. Trust Put to the Test
2. a Testcase…
1. **… for a Cognitive Trust Model**
4. Implementation
5. Conclusions & further work

Trust models

- C. Castelfranchi, R. Falcone, and E. Lorini, “A non reductionist approach to trust,” in *Computing with Social Trust and Reputation*, ser. Human-Computer Interaction Series, J. Golbeck, Ed. Springer, 2008, pp. 45-72.
- A. Herzig, E. Lorini, J. F. Hübner, and L. Vercouter, “A logic of trust and reputation,” *Logic Journal of the IGPL*, vol. 18, no. 1, pp. 214-244, 2010.
- A. J. I. Jones, “On the concept of trust,” *Decis. Support Syst.*, vol. 33, no. 3, pp. 225-232, 2002.
- S. Marsh and P. Briggs, *Computing with social trust*. Springer, 2009, ch. Examining trust, forgiveness and regret as computational concepts.
- …

A cognitive trust model

- A. Herzig, E. Lorini, J. F. Hübner, and L. Vercouter, “A logic of trust and reputation,” *Logic Journal of the IGPL*, vol. 18, no. 1, pp. 214-244, 2010.

Based on cognitive (sociological) theories

➔ Models trust and reputation as they are found in the human society

➔ Realistic model

Trust/Reputation: ingredients

Reputation
- in a group *I*
- of an agent *j* gent *j*

- w.r.t. an action α
- in order to achieve a goal φ
- according to the circumstances κ

Trust/Reputation in action

Agent *j* has reputation in group *l* to do α w.r.t. φ in the circumstances κ iff

- *i* has the potential goal φ in the circumstances κ
- *i* believes that from now on, if it has the goal φ and κ holds, then
  - *j* will be capable to do α
  - *j*, by doing α, will ensure φ to be true at some point
  - *j* will intend to do α

Trust/Reputation in inaction

Agent *j* has reput*ot* to do α w.r.t. φ in the circumstances κ iff

Our contribution

- *I* has the potential group goal φ in the circumstances κ
- it is public for the group *I* that from now on, if the group *I* has the goal φ and κ holds, then
  - *j* will be capable to do α
  - *j*, by doing α, will ensure φ to be *always false*
  - *j does not* intend to do α

Outline

3. Trust Put to the Test
2. **a Testcase…**
1. … for a Cognitive Trust Model
4. Implementation
5. Conclusions & further work

A Wikipedia-based testcase

Patrol: Set of Wikipedia community members helping editors maintain reasonable quality

A number of tools is available to Recent Changes Patrollers (RCPs)

- None allows for knowledge transfer among them

Needy edit: Edit requiring improvement in some manner

Bad edit: Edit that may need to be entirely removed

Good edit: Edit that is neither needy nor bad

Trust/Reputation: ingredients

Trust
- of an agent *i*
- towards an agent *j*
- w.r.t. an action α
- in order to achieve a goal φ
- according to the circumstances κ

The RC

The Wikipedia
1. Good edits
2. Bad edits

the integrity
of Wikipedia articles

All possible circumstances

Trust/Reputation: ingredients

Reputation
- in a group *I*
- of an agent *i*
- w.r.t. an ac
- in order t
- according

The Wikipedia contributor

- The set of all patrollers?
  - Most patrollers do not have any opinion about most contributors
- The set of patrollers having a well-founded opinion of the contributor

Trust/Reputation in action

Agent *i* is disposed to trust agent *j* to do α w.r.t. φ in the circumstances κ iff

- *i* has the potential goal φ in the circumstances κ
- *i* believes [illegible] goal φ and κ holds, then

True

We only have to check that

  - *j* will be capable to do α
  - *j*, by doing α, will ensure φ to be true at some point
  - *j* will intend to do α

Similarly for reputation in action and trust/reputation in inaction

We only have to estimate if (I)

(w.r.t. doing a good edit)

- a given Wikipedia contributor is capa… do a good edit — True
- by doing a good edit, the contributor will ensure the article to be consistent at some point
- the contributor intends to do a good… — True

(w.r.t. refraining from doing a bad edit)

- a given Wikipedia contributor is capa… do a bad edit — False
- by doing a bad edit, the contributor will ensure the article to be always inconsistent
- the contributor does not intend to do a bad edit

We only have to estimate if (II)

(the RCP believes or it is public in the given set of RCPs that)

- a given Wikipedia contributor is capable and intends to do a good edit
- she does not intend to do a bad edit

Use statistics about the contributor's previous activity

- The number of good edits is above a given threshold $\Rightarrow$ she can be assumed to be capable and to intend to do a good edit
- (similarly for bad edits)

We only have to estimate if (III)

- An edit is good or bad

Use heuristics

- Their effectiveness must be evaluated
- (cf. further work)

Enable RCPs to provide feedback

Outline

3. **Trust Put to the Test**
2. a Testcase…
1. … for a Cognitive Trust Model
4. Implementation
5. Conclusions & further work

Experimental setup

Data have been crawled

- from the English Wikipedia
- throughout 24 hours
- starting from Fri Oct 22 16:45:49 CEST 2010

183386 edits performed on 103811 Wikipedia pages by 30806 contributors

- 103146 edits (56%) involved Wikipedia pages other than articles
  - administrative pages, templates, talks…
- 28528 edits (16%) were performed by bots
  - (semi-)automated tools carrying out maintenance tasks

Experimental methodology

3825 edits performed by humans on articles have been manually reviewed

- `good`, `needy`, `bad`, `possibly bad`, `unknown`

Predictions of our algorithm have been compared to the manual assessments

- Only contributors who performed ≥2 non-`unknown` edits were considered
  - 2944 edits by 398 contributors

Experimental results

4-5% false positives
• good edits considered as bad

86-87% precision

actual / expected	bad	needy	good
bad	9%	0%	4%
needy	2%	0%	2%
good	5%	0%	77%

(a) *possibly bad* edits considered as *good*

actual / expected	bad	needy	good
bad	11%	0%	5%
needy	2%	0%	2%
good	5%	0%	76%

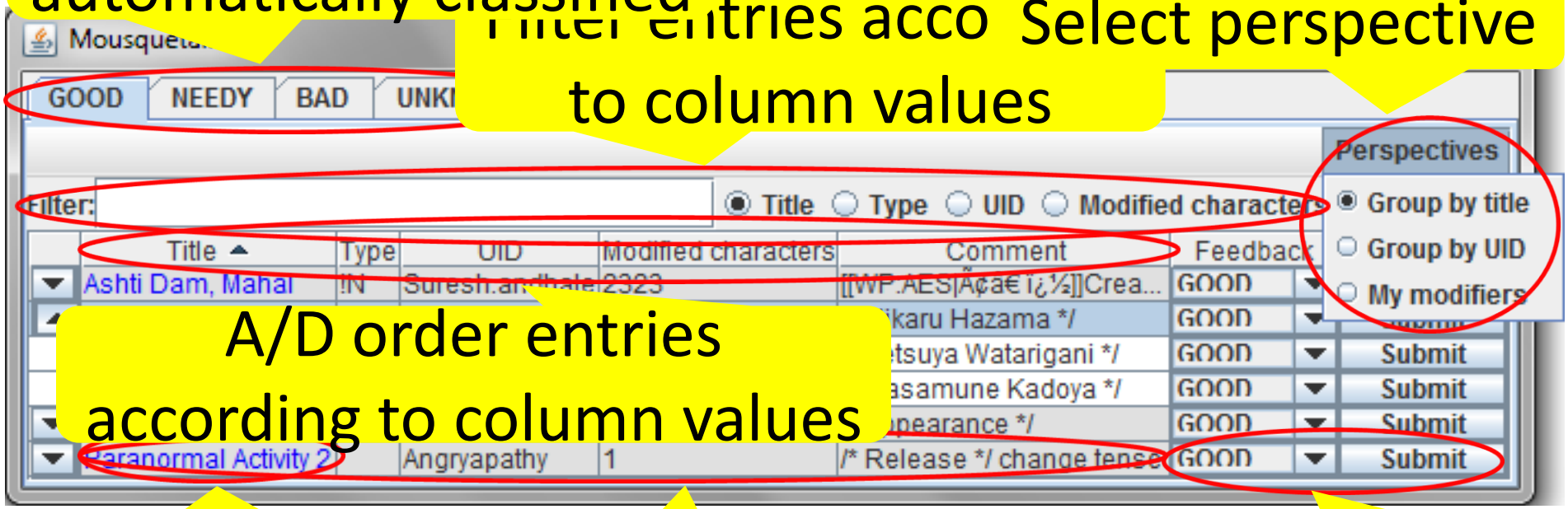
(b) *possibly bad* edits are considered as *bad*

False negatives
• bad edits considered as good

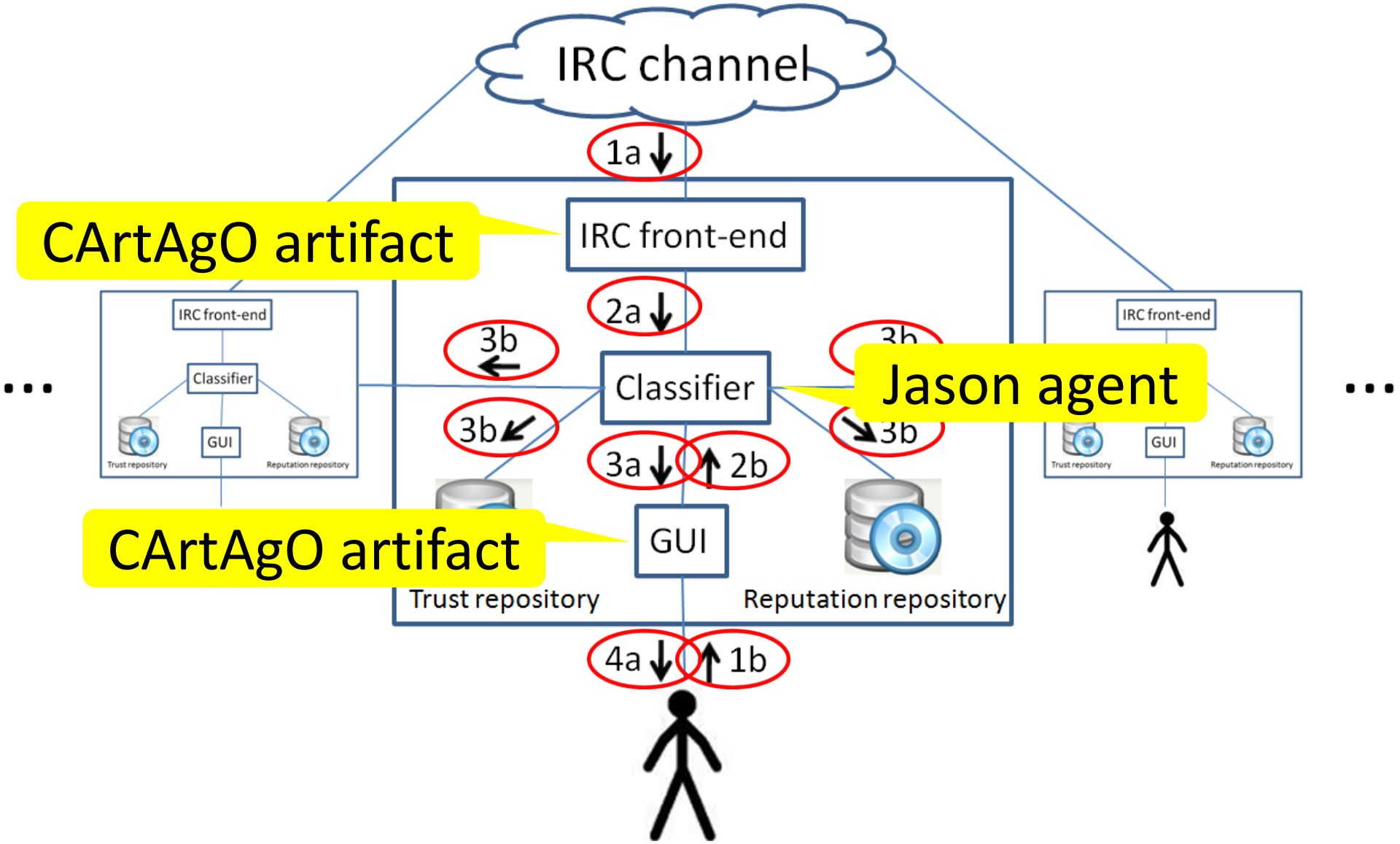
Outline

3. Trust Put to the Test
2. a Testcase…
1. … for a Cognitive Trust Model
4. **Implementation**
5. Conclusions & further work

MOUSQUETAIRE'S GUI



Behind the scenes



Outline

3. Trust Put to the Test
2. a Testcase…
1. … for a Cognitive Trust Model
4. Implementation
5. **Conclusions & further work**

Conclusions

- We extended Herzig et al.'s conceptual framework
- We presented an instantiation of such a framework
- We implemented and evaluated such instantiation

Further work

- Defining effective heuristics to overcome the cold-start problem
- Making our implementation available to the Wikipedia community
  - Current state available at `http://labh-curien.univ-st-etienne.fr/~decoi/Mousquetaire.zip`