A Computational Model of Trust Based on Message Content and Source

Célia da Costa Pereira, Andrea G. B. Tettamanzi, Serena Villata

A Computational Model of Trust Based on Message Content and Source

(Extended Abstract)

Célia da Costa Pereira
Univ. Nice Sophia Antipolis
I3S, UMR 7271, France
celia.pereira@unice.fr

Andrea G. B. Tettamanzi
Univ. Nice Sophia Antipolis
I3S, UMR 7271, France
andrea.tettamanzi@unice.fr

Serena Villata
Inria Sophia Antipolis
I3S, UMR 7271, France
serena.villata@inria.fr

ABSTRACT

We propose a general possibilistic framework to determine the agent's trust degree in a source, starting from the content of the messages such source provides and based on the beliefs of the agent about the capability of the source to provide "useful information". The result is a framework with unique characteristics, which combines experience-, reputation-, content-, and category-based models of trust in one coherent computational model.

Categories and Subject Descriptors

I.2.11 [**Distributed Artificial Intelligence**]: Multiagent systems

Keywords

Trust; Distrust; BDI Agents; Argumentation Theory; Goals

1. INTRODUCTION

The extent to which new information is accepted by an agent directly depends on the content of the new claim and on how much the agent trusts the source providing it. However, trust may also be influenced by information content. Indeed, even tough I might not particularly trust a source, if it provides me a claim which is consistent with my beliefs, I will not change my beliefs. However, my degree of trust for that source may increase. Trust depends thus on the agent's own beliefs in general and, in particular, on the agent's opinion about the capability of the source to convey useful information. In real-world situations, an agent's beliefs about a source may be incomplete, for they may derive from the opinions of other (partially) known agents and the agent may have had few (or none) exchanges with the source.

On the other hand, only an agent endowed with goals and beliefs can trust another agent [2]. In other words, if an agent needs to trust a source, it is because it needs "something" from that source that could help it fulfill its own goals. Therefore, the agent's beliefs about the source's goals in comparison with its own goals must also play an important role in computing trust. These beliefs can be constructed from the agent's past interactions and the source's reputation and/or recommendations.

Because we are aware that trust is not always the complement of distrust, here, we consider the bipolar side of trust. Our key idea here is to capture the fact that some pieces of information can contribute to increase or decrease trust, and other pieces of information can contribute to increase or decrease distrust.

We propose a general possibilistic and argumentation-based formalism which combines the above-mentioned components to compute the agent's trust/distrust in a source. The aim of our proposal is twofold. First, to provide a computational trust model able to cope with the uncertainty due to the agent's knowledge about a source being possibly incomplete. We use possibility theory to this aim. Second, to capture the internal process followed by the agent to decide which information sources are reliable, based on how convincingly they make their claims. We adopt an argumentation framework to cope with this second aim.

Appears in: *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2015), Bordini, Elkind, Weiss, Yolum (eds.), May 4–8, 2015, Istanbul, Turkey.*

2. THE PROPOSED TRUST MODEL

We build our model of trust on top of the BDI framework proposed in [3], which uses argumentation [1] both to resolve conflicts among trust-weighted information sources and to revise beliefs and possibility theory [4] to represent graded beliefs and reason about the world. We refer the reader to the cited literature for basic definitions about possibility theory, argumentation, and our underlying BDI framework.

Some pieces of information contribute to increase or decrease the trust that an agent has in a source, and others contribute to increase or decrease distrust. This is why, like [6], we suppose that trust and distrust are separated—are not the opposite ends of a single continuum—but linked dimensions that can coexist and have different antecedents and consequences.

Like [2], we consider trust as beliefs: an agent trusts a source s if and only if it somehow believes that s will be able to help it improve the satisfaction degree of its goals. We will also define distrust as a belief of an agent: an agent distrusts a source s if and only if it somehow believes that s will try to prevent it to reach its goals. We should always keep in mind that trust and distrust in s are to be construed conceptually as if they were defined as follows:

$$\mathrm{trust}(s) \equiv \mathbf{B}(\text{“}s \text{ is trustworthy”}), \quad (1)$$

$$\mathrm{distrust}(s) \equiv \mathbf{B}(\text{“}s \text{ is untrustworthy”}),. \quad (2)$$

Notice that proposition "s is untrustworthy" is the logical negation of "s is trustworthy".

A straightforward consequence of bipolar Eq. 1 and 2 is that trust and distrust are connected by the following mutual constraints:

$$\mathrm{trust}(s) > 0 \Rightarrow \mathrm{distrust}(s) = 0, \quad (3)$$

$$\mathrm{distrust}(s) > 0 \Rightarrow \mathrm{trust}(s) = 0. \quad (4)$$

In case of total ignorance, $\mathrm{trust}(s) = \mathrm{distrust}(s) = 0$.

Our proposal is grounded on the fact that when a new message arrives from a source s an agent will:

- compute a trust value τ' and an associated distrust value δ', which, together, mirror how convincingly the source has argued for its claims up to that moment; these are based on its consistency (inconsistency) with the agent's current beliefs, which depend on all the previous arguments A provided by s and by other sources:

$$\tau' = \max\{0, \min_{A:s\in\mathrm{src}(A)} \alpha(A) + \max_{A:s\in\mathrm{src}(A)} \alpha(A) - 1\}, \quad (5)$$

$$\delta' = \max\{0, 1 - \max_{A:s\in\mathrm{src}(A)} \alpha(A) - \min_{A:s\in\mathrm{src}(A)} \alpha(A)\}; \quad (6)$$

  where $\alpha(A)$ represents the acceptability degree of argument A computed using the fuzzy labeling algorithm of [3].

- consider a set of categories used to compute a trust value τ_{cat} and the associated distrust value δ_{cat}, which take into account the category of the source (whether it is rational or not, malicious or benign, etc.);

- compute a third value of trust τ_{goal}, together with the associated value of distrust δ_{goal}, which take into account the overlap (or not) between the agent's goals and the source's goals: the beliefs about the source's goals are compared with the agent's own goals; four cases are distinguished: (a) if s reaches its goals then the agent does too, (b) the two sets of goals are independent, (c) there is an overt conflict between the two sets of goals, and (d) if the agent reaches its goals then the source does too.

- combine $(\tau_{\mathrm{cat}}, \delta_{\mathrm{cat}})$ and $(\tau_{\mathrm{goal}}, \delta_{\mathrm{goal}})$ into a trust value τ'' and the associate distrust value δ'' using Kleene-Dienes fuzzy implication [5], so that the overall trust/distrust in source s depends on the uncertainty about the sources real category and the uncertainty about the source's goals;

- finally, combine (τ', δ') and (τ'', δ'') into an overall value τ of trust and δ of distrust to be used to decide the membership degree of the argument in $\mathcal{A}$:

$$\tau = \min\{\tau', \tau''\}, \quad (7)$$

$$\delta = \min\{\delta', \delta''\}. \quad (8)$$

  The choice of $\min$ as the aggregation operator is motivated by the fact that an agent should believe source s is (un)trustworthy if it believes it is (un)trustworthy because of the past arguments it provided *and* of its goals and category.

To summarize, an agent receives a new argument A from a source with trust τ. This argument becomes part of the fuzzy set of the trusted arguments. A fuzzy labeling algorithm is then run in order to compute a new labeling for all arguments. These new values of acceptability of the arguments are then used to update the agent's belief degrees. This results in the BDI agent architecture schematically illustrated in Figure 1. An agent interacts with the world by receiving arguments A from one or more *sources*. Special cases of arguments are trust-related arguments A_{cat} reporting about the category of other agents. More precisely, they represent the recommendation and reputation-based information the agent receives from sources about other sources as well as the experience-based information about the reliability of sources that provided arguments on the basis of which past actions were planned.

The agent's internal mental state is described by a fuzzy set of trustful arguments $\mathcal{A}$, from which the beliefs of the agent are derived. The Trust module, which is the core of this paper, assigns

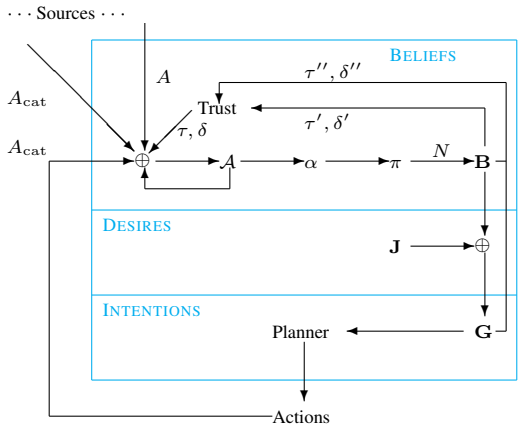

Figure 1: A schematic illustration of the BDI agent architecture within which the proposed model of trust is set.

a trust degree τ and a distrust degree δ to each source based on both the agent's own goals and its beliefs about the category and the goals of that source. These beliefs, and thus the trust and distrust degrees associated to the sources are updated every time $\mathcal{A}$ changes. As new arguments are received, they are added to $\mathcal{A}$ with the same membership degree as the degree τ to which their most trustworthy source is trusted. A fuzzy labelling algorithms [3] computes the degree α to which every argument is accepted. From α, a possibility distribution π is computed, from which an explicit representation of the agent's beliefs $\mathbf{B}$ is constructed as the necessity measure N of π. The beliefs, together with justified desires $\mathbf{J}$ of the agent allow to generate the goals $\mathbf{G}$. The agent then plans its actions to achieve the selected goals by means of a planner module. The results of the agent's actions are used to construct the set of experience-based arguments about the trustworthiness of the sources which provided arguments relevant to the plan.

3. CONCLUSION

We have developed a general possibilistic framework for trust computation in BDI agents, which combines both experience-based and reputation-based trust (expressed under the form of arguments), goal-based trust, as well as information-content-based trust to support the agent in assigning a trust/distrust degree to sources.

REFERENCES

[1] P. Besnard and A. Hunter. *Elements of Argumentation*. The MIT Press, 2008.

[2] C. Castelfranchi and R. Falcone. Social trust: A cognitive approach. In C. Castelfranchi and Y.-H. Tan, editors, *Trust and Deception in Virtual Societies*, pages 55–90. Springer, 2001.

[3] C. da Costa Pereira, A. Tettamanzi, and S. Villata. Changing one's mind: Erase or rewind? In *IJCAI*, pages 164–171, 2011.

[4] D. Dubois and H. Prade. *Possibility Theory*. Plenum, New York, 1988.

[5] M. Mas, M. Monserrat, J. Torrens, and E. Trillas. A survey on fuzzy implication functions. *Trans. Fuz Sys.*, 15(6):1107–1121, 2007.

[6] D. H. McKnight and N. L. Chervany. Trust and distrust definitions: One bite at a time. In *Trust in Cyber-societies*, volume 2246 of *Lecture Notes in Computer Science*, pages 27–4. Springer, 2000.