Trust Model for Human Agent Interaction

Nuno Xu
nuno.xu@tecnico.ulisboa.pt

Supervisors: Rui Prada and Ana Paiva
Técnico Lisboa (Taguspark)
Universidade de Lisboa
Av. Prof. Dr. Aníbal Cavaco Silva
Porto Salvo, Portugal
http://tecnico.ulisboa.pt/en/

1 Introduction

Cognitive computational trust modelling started from the classical introduction from Castlefranchi and Falcone [1], which established the initial view of representing trust as a collection of beliefs and intentions about delegating a certain task to the trustee. It also provided the vision that trust is an emergent concept mainly composed by our belief in the trustor's ability for the task in question, and it's willingness to perform the task in question. In recent years some attempts have been made to improve upon this aspects [2–8], further addressing issues like reputation [7], and belief representation [2, 6].

For the purpose of this thesis, we sought out to create a model that showcased the multidimensionality of trust and how much do the different features affect a requested trust evaluation.

2 Trust Model

In an effort to create a working trust model iteratively, we will start by simplifying the model described by Castelfranchi and Falcone [1], by removing the effects of outside influence in Trust. We also do not take into account long term considerations of the trustor's goal, reducing contextual scope to just the task being performed by the trustor. So Trust is represented by a 3-tuple:

- The trustor ($\mathbf{X}$);
- The trustee ($\mathbf{Y}$);
- A task ($\boldsymbol{\tau}$) defined by the pair (α, ρ), where $\boldsymbol{\alpha}$ is the action entrusted to the trustee, that possibly produces an outcome $\boldsymbol{\rho}$, contained in the goal of $\mathbf{X}$.

$$TRUST(X\ Y\ \tau) \tag{1}$$

We seek to represent the trustee as a set of features from which an overall evaluation may be retrieved. These features provide a representation of the

trustee's abilities in various fields, as well as concerns related to willingness, such as task preference. So a trustee Y's feature set S_y can be as seen in 2.

$$S_y = \{cooking, writing, preference\ to\ cook\} \tag{2}$$

These features must be able to provide 2 values:

- trust - a value for how much trust we have in this trustee's specific feature;
- certainty - the degree of how much we believe this trust assumption to be true, tanking into account factors like how many times, or how long ago did we last affirm this belief.

The trust and certainty values are provided by a collection of *belief sources* contained in each feature, representing the different pieces of information that contribute to form our belief in said feature. This *belief sources* can be divided in 3 kinds:

- Bias - a source representing
- certainty - the degree of how much we believe this trust assumption to be true, tanking into account factors like how many times, or how long ago did we last affirm this belief.

This makes it easier to visualize a certain trustee's trust reasoning, as we can observe all the factors that contribute towards an evaluation. The specificity of the concrete features is purposefully left generic in order make the model fit in different scenarios. So an ontology must be provided for the model to have a collection of features that can be assigned to trustees.

The task is composed by a set of related features with a given weight. This represents the features most closely related to the task at hand, and their importance to he completion of the task.

2.1 Implementation

Implementation wise, the model will be first implemented by using a simple class structure, as seen in figure 2. In this diagram, the main actor is the Agent, which contains a list of Trustees with features that the Agent has been able to perceive from received sources. For now the sources must be given by the simulation environment, but an interpreter should be implemented to sort out and transform the perceptions received by the environment into sources for belief features. A simple simulation example can be performed as following:

1. Instantiate Agent A(nna) and Agent B(ob);
2. Instantiate Trustee B and assign as A's trustee;
3. Insert Direct Contact Source with Feature ID "Cooking"; this should create a new Trust Feature in Trustee B;
4. Calculate Trust with task "Cook";

3 Example Case

4 Conclusion

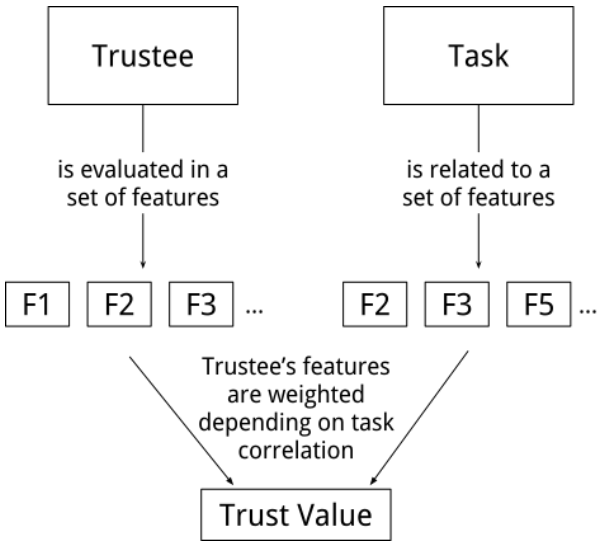

Figure 1: Trustee Features Representation

References

1. Castelfranchi, C., Falcone, R.: Principles of trust for MAS: cognitive anatomy, social importance, and quantification. Proceedings of the International Conference on Multi Agent Systems (1998) 72–79
2. Piunti, M., Venanzi, M., Falcone, R.: Multimodal Trust Formation with Uninformed Cognitive Maps (UnCM) (Extended Abstract). Aamas (2012) 1241–1242
3. Sutcliffe, A., Wang, D.: Computational Modelling of Trust and Social Relationships. Journal of Artificial Societies and Social Simulation **15**(1) (aug 2012) 523–531
4. Singh, M.: Trust as dependence: a logical approach. The 10th International Conference on Autonomous ... (2011) 863–870
5. Castelfranchi, C., Falcone, R.: Trust Theory. 1 edn. John Wiley & Sons, Ltd, Chichester, UK (mar 2010)
6. Pinyol, I.: Reputation-Based Decisions for Cognitive Agents (Thesis Abstract). Doctoral Mentoring Program (Aamas) (2009) 33
7. Sabater, J., Paolucci, M., Conte, R.: Repage: REPutation and ImAGE among limited autonomous partners. Jasss **9**(2) (2006) 117–134
8. Neville, B., Pitt, J.: A Computational Framework for Social Agents in Agent Mediated E-commerce. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Volume 3071. (2004) 376–391

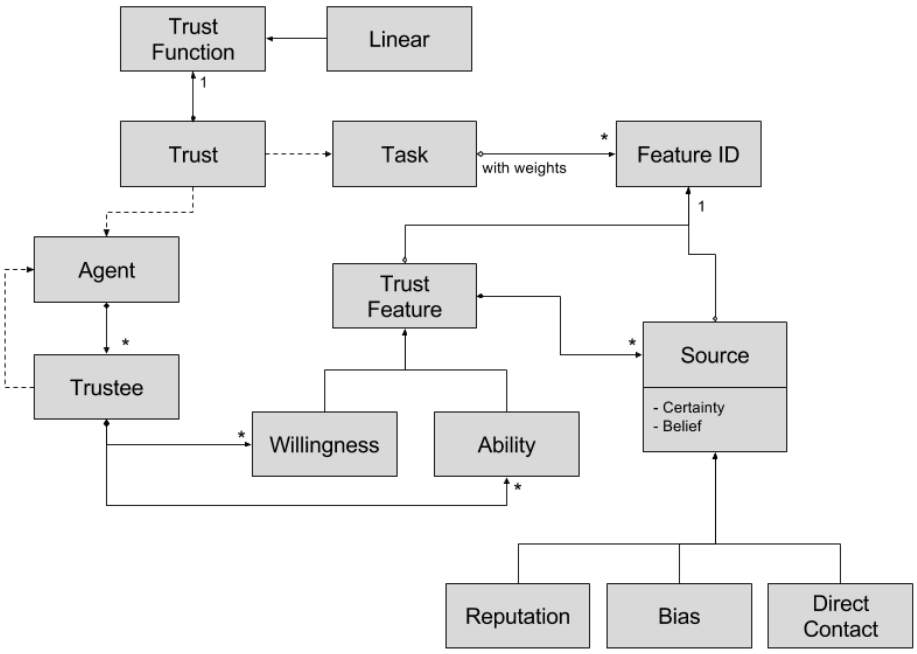

Figure 2: Class Diagram