The Semantic Interpretation of Trust in Multiagent Interactions

Anup K. Kalia
Department of Computer Science
North Carolina State University
Raleigh, NC 27695-8206, USA
akkalia@ncsu.edu

Introduction

In a multiagent system, a *truster* keeps estimating and revising its trust for a *trustee* based on its interactions with *trustee*. Understanding these interactions and estimating trust from them is an interesting and challenging topic and therefore, several works have been pursued in this direction. Scissors et al. (Scissors et al. 2009) explore the linguistic similarity in chat messages to estimate trust between message senders and receivers. Adalı et al. (Adalı, Sisenda, and Magdon-Ismail 2012) calculate the relationship strength between two users in *Twitter* based on their different social and behavioral aspects. (DuBois, Golbeck, and Srinivasan 2011) provide an algorithm to compute trust and distrust in a social network. The above approaches are promising but they are limited to numerical heuristics. Such approaches are justifiably criticized for missing the essential intuitive considerations of trust, e.g., regarding the autonomy of the participants and the vulnerability of the truster to decisions by the trustee. The richer approaches, however, have not lent themselves well to computational techniques that could be applied in practice (Castelfranchi and Falcone 2010). Therefore, we seek to bridge the above gap between theory and practice. Specifically, we propose a computational model of trust founded on commitments that supports agents determining their trust for others based on their interactions.

A commitment describes a normative relationship between two agents giving a high-level description of what one expects of the other. We provide a simple probabilistic model for estimating trust based on the outcomes of commitments. Importantly, not only can commitments be inferred from structured interactions between people, they can also be determined automatically from text-based interactions (email or chat) (Kalia et al. 2013). We provide a method to train the parameters of our model based on data regarding judgments of trust given commitment outcomes. Our method can thus potentially help compute in a user-specific manner how much a user would trust another party given their mutual interactions.

Copyright © 2014, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Intuition on Trust and Commitment

A commitment C(*debtor, creditor, antecedent, consequent*) means that the debtor commits to bringing about the consequent for the creditor provided the antecedent holds. For example, C(*Buck, Selia, deliver, pay*) means that Buck (buyer) commits to Selia (seller) to paying a specified amount provided Selia delivers the goods. When Selia delivers, the commitment is detached. When Buck pays, the commitment is discharged or satisfied. If Selia delivers but Buck does not pay, the commitment is violated. The above example suggests how commitments and trust relate. If Buck discharges the commitment, it brings a positive experience to Selia and Selia's trust for Buck may increase; if Buck violates the commitment, it brings a negative experience to Selia and Selia's trust for Buck may decrease. Figure 1 illustrates the intuition graphically.

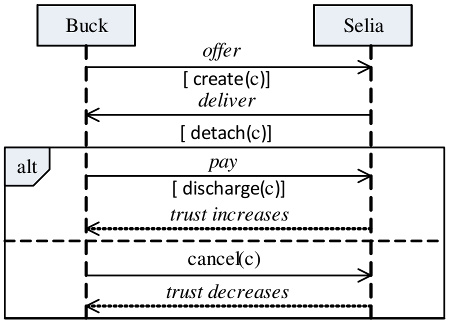

Figure 1: Trust updates based on a commitment progression.

Updating Trust based on Commitments

We adopt Wang and Singh's (2010) trust model, which represents trust as evidence $\langle r, s\rangle$. Here, $r \geq 0$ and $s \geq 0$ respectively represent the positive and negative experiences the truster has with the trustee. Both r and s are real numbers. Wang and Singh calculate trust as the probability of a positive outcome as $\alpha = \frac{r}{r+s}$. Suppose Buck and Selia transact 10 times and exactly eight transactions succeed from Selia's perspective. Then Selia's trust in Buck would be 0.8.

The basic idea is for each truster to maintain evidence $\langle r, s\rangle$ about each trustee. The initial evidence, $\langle r_{in}, s_{in}\rangle$, represents the truster's bias. An interaction may yield a positive, negative, or a neutral outcome. In these cases, the evidence is updated by respectively adding $\langle i_r, 0\rangle$, $\langle 0, i_s\rangle$, and $\langle \lambda i_r, (1-\lambda)i_s\rangle$, where $\lambda \in [0, 1]$. In essence, we characterize each truster via five parameters (r_{in}, s_{in}, i_r, i_s, λ).

Evaluation

We evaluated our approach via an empirical study with 30 subjects (computer science students). We asked the subjects to read 33 emails selected from the Enron email corpus (Fiore and Heer 2004) and estimate a trust value ranging from 0 to 1 between the senders and receivers of email. The emails were selected on the basis of containing sentences that indicate commitment creation, satisfaction, or violation. The sentences indicating commitments were identified using Kalia et al.'s (2013) approach. We augmented the dataset with 28 synthetic sentences indicating commitment satisfaction or violation, which do not occur frequently in the corpus. We collected the trust values (actual) from the subjects from the emails assigned to them. We divided the data collected from subjects into three-fold training and test data and learned trust parameters for each subject (r_{in}, s_{in}, i_r, i_s, λ) that minimize the mean absolute error (MAE) between predicted and actual trust values (Kalia, Zhang, and Singh 2013). We performed more evaluations to find whether MAEs obtained can be reduced further by considering a discount window and the weight of commitments created in agents' interactions. A discount window is based on the most recent experiences perceived by agents from their interactions. The weight of a commitment is calculate based on different features present in it such as modal verbs, deadlines, and so on.

Future Work

We are exploring other methods to predict trust values apart from commitments. For a start, we have conducted following investigations; (1) whether the outcome of a goal affects trust more than the outcome of a commitment and (2) whether an agent's mood affects its trust toward satisfying a commitment. To evaluate our hypotheses we performed an empirical evaluation with 30 subjects where we asked them to play a variant of Gal et al.'s (2010) Colored Trails game. Our variant provides a chat interface, through which subjects negotiate and exchange tiles and express emotions toward opponents. During the game subjects recorded their interactions with their opponents and filled a feedback form with their trust and mood before and after each round in a game. From their interactions, we manually identified commitments and emotions expressed by the subjects. We used the data collected to train different Bayesian models constructed based on our assumptions using Expectation Maximization. From the preliminary results, we found that a commitment outcome affects trust more than goal outcome and both mood and trust are important in predicting the outcome of a commitment instead of considering either trust or mood. We also observed that trust affects mood more than mood affecting trust.

We plan to improve the current version of the game to include virtual agents programmed with different strategies. Later, we will hire subjects from Amazon Turk to play games against these agents. From their games, we will collect their interactions with agents and their trust, moods, and emotions toward them. The primary reasons to create virtual agents is to collect a larger dataset than our previous experiments and manipulate and evaluate subjects' behaviors based on different strategies set for agents. This may also ensure more truthful trust estimation behavior from subjects compared to estimating trust from emails.

References

Adalı, S.; Sisenda, F.; and Magdon-Ismail, M. 2012. Actions speak as loud as words: Predicting relationships from social behavior data. In *Proceedings of the 21st International Conference on World Wide Web*, WWW, 689–698. ACM.

Castelfranchi, C., and Falcone, R. 2010. *Trust Theory: A Socio-Cognitive and Computational Model*. Agent Technology. Chichester, UK: John Wiley & Sons.

DuBois, T.; Golbeck, J.; and Srinivasan, A. 2011. Predicting trust and distrust in social networks. In *Proceedings of 3rd International Conference on Social Computing*, 418–424.

Fiore, A., and Heer, J. 2004. UC Berkeley Enron email analysis.

Gal, Y.; Grosz, B.; Kraus, S.; Pfeffer, A.; and Shieber, S. 2010. Agent decision-making in open-mixed networks. *Artificial Intelligence* 174(18):1460–1480.

Kalia, A.; Nezhad, H. R. M.; Bartolini, C.; and Singh, M. P. 2013. Monitoring commitments in people-driven service engagements. In *Proceedings of the 10th IEEE Conference on Services Computing*, 1–8.

Kalia, A. K.; Zhang, Z.; and Singh, M. P. 2013. Trustworthy decision making via commitments. In *Proceedings of the 15th AAMAS Workshop on Trust in Agent Societies (Trust)*, 24–35.

Scissors, L. E.; Gill, A. J.; Geraghty, K.; and Gergle, D. 2009. In CMC we trust: The role of similarity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI, 527–536. ACM.

Wang, Y., and Singh, M. P. 2010. Evidence-based trust: A mathematical model geared for multiagent systems. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)* 5(4):14:1–14:28.