ECAI 2014
T. Schaub et al. (Eds.)

doi:10.3233/978-1-61499-419-0-1043

Estimating Trust from Agents' Interactions via Commitments

Anup K. Kalia[1] and **Zhe Zhang**[2] and **Munindar P. Singh**[3]

Abstract. How an agent trusts another naturally depends on the outcomes of their interactions. Previous approaches have treated the outcomes in a domain-specific way. We propose an approach relating trust to the domain-independent notion of commitments. We conduct an empirical study to evaluate our approach, in which subjects read emails extracted from the Enron dataset (augmented with some synthetic emails for completeness), and estimate trust between each pair of communicating participants. We propose a probabilistic model for trust based on commitment outcomes and show how to train its parameters for each subject based on the subject's trust assessments. The results are promising, though imperfect. Our main contribution is to launch a research program into computing trust based on a semantically well-founded account of agent interactions.

1 Introduction

Understanding multiagent interactions and estimating trust from them is an interesting and challenging topic. Several approaches [7, 1] have been proposed to estimate trust from interactions. However, they are limited to numerical heuristics and ignore the essential intuitive aspects of trust. In contrast, we propose a probabilistic model of trust based on commitment outcomes that supports agents to determine their trust for others based on their interactions. Our model captures the intuition that both the truster and the trustee are autonomous and the truster is vulnerable to decisions of the trustee [2].

Commitments are important for trust because they can be identified from agents' interactions and can help us characterize the outcomes of such interactions in high-level terms. 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(*Bob, Alice, deliver, pay*) means that Bob (buyer) commits to Alice (seller) to paying a specified amount provided Alice delivers the goods. When Alice delivers, the commitment is detached. When Bob pays, the commitment is discharged or satisfied. If Alice delivers but Bob does not pay, the commitment is violated. In essence, a commitment describes a social relationship between two agents giving a high-level description of what one agent expects of the other. As a result, it is natural that commitments (and their satisfaction or violation) be used as bases for trust. In the above example, if Bob discharges the commitment, it brings a positive experience to Alice and Alice's trust for Bob may increase; if Bob violates the commitment, it brings a negative experience to Alice and Alice's trust for Bob may decrease.

We conduct an empirical evaluation on emails automatically analyzed using our previous approach [5]. We show how to train the model parameters so as to capture a user model indicating each user's propensity to trust given commitment outcomes. Our evaluations yield promising, but imperfect, results on the viability of inferring trust from the commitments arising in interactions, suggesting the need for better extraction techniques. Our main contribution is to show how trust can be computed, not just theorized about, via the domain-independent concept of commitments.

2 Model of Trust based on Commitments

We adopt Wang and Singh's [8] 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 experience. 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 (i_r, i_s, r_{in}, s_{in}, λ).

2.1 Learning Trusters' Trust Parameters

Trust assessment is subjective. Trusters differ in how they update their trust for a trustee when a commitment is discharged or violated, respectively. Therefore, we learn a specific truster's parameters based on positive, negative, and neutral experiences with various trustees and the truster's actual trust in them. For the k^{th} trustee, let α_k represent the truster's actual (as revealed) and $\hat{\alpha_k}$ the truster's predicted trust in k. Let E_k^+, E_k^-, and E_k represent the numbers of positive, negative, and neutral experiences, respectively. Then,

$$\hat{\alpha_k} = \frac{r_{in} + \mathrm{E}_k^+ i_r + \lambda \cdot \mathrm{E}_k i_r}{r_{in} + s_{in} + \mathrm{E}_k^+ i_r + \mathrm{E}_k^- i_s + \mathrm{E}_k(\lambda i_r + (1-\lambda) i_s)} \qquad (1)$$

Via nonlinear least-squares regression technique that uses trust region reflective algorithm [3], we estimate the truster's parameters to minimize the mean absolute error (MAE) of prediction, $\sum_{k=1}^{n} |\hat{\alpha_k} - \alpha_k|$.

[1] NC State University, Raleigh, US, email: akkalia@ncsu.edu
[2] NC State University, Raleigh, US, email: zzhang13@ncsu.edu
[3] NC State University, Raleigh, US, email: mpsingh@ncsu.edu

We now present our hypothesis, i.e., the above approach to predict trust values by *learning trust parameters* for each subject is more accurate than using *fixed trust parameters* for all the subjects.

3 Evaluation and Results

We evaluated our approach via an empirical study with 30 subjects (computer science students). The subjects read 33 emails selected from the Enron email corpus [4, 6] and provided 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—such sentences were identified using Kalia et al.'s [5] 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 from the subjects from the emails assigned to them. We conducted three-fold evaluation wherein we learned trust parameters for each subject (r_{in}, s_{in}, i_r, i_s, λ) that minimize MAE between predicted and actual trust values. For verifying our hypothesis, we calculated the MAE for λ in intervals of 0.1 to 0.9. Then, we calculated the MAE by learning the λ (L(λ)) itself. Based on the above MAEs, we obtained a customized λ (fixed or learned) for each subject. A customized λ for a subject refers to the value of λ for which the MAE is minimum. We represent the MAEs obtained using customized λs for all subjects as C(λ) in Figure 1. Finally, we arbitrarily selected some fixed configurations of parameters (F1 = $\langle 1, 1, 1, 1, 0.5\rangle$, F2 = $\langle 2, 1, 1, 1, 0.5\rangle$, F3 = $\langle 1, 2, 1, 1, 0.5\rangle$). F1 indicates no bias in the initial trust perception whereas F2 and F3 indicate positive and negative biases respectively. And, λ=0.5 in the fixed configurations indicates equal trust increments for the neutral experiences. From the results, we observe that the median of C(λ) (0.162) is less than the medians of all other approaches. However, from the one-tailed t-test, we found that the mean of C(λ) is not significantly lower than the means of other approaches. The overall results suggest that although hypothesis is supported on descriptive grounds (apparent differences in MAE distributions), it cannot be asserted based on statistical significance tests.

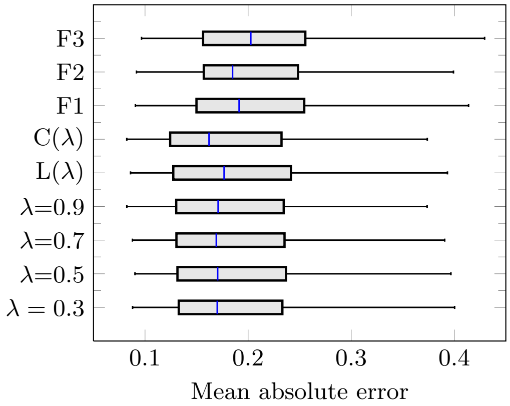

Figure 1. MAE for predicting trust values.

4 Discussion and Future Work

The main contribution of our approach is to develop a computational approach for trust that overlays a domain-independent concept describing the social relationships and outcomes of interactions between agents. We evaluated our approach on an email dataset, comparing the means of the MAEs. The results indicate that our approach yields a correlation between subjects' intuitions regarding trust values and those computationally predicted values. The limitation of our result may be due to the following reasons: (1) lack of adequate data (2) a greater fraction of experiences being judged neutral than positive or negative or (3) too small of a sample size for obtaining a sufficiently low p-value. Also, we lack an existing approach with which to compare our results.

In the future, we plan to address these limitations by adopting an incentive scheme that motivates subjects to provide trust values truthfully. Moreover, there is no reason to be limited to commitments: indeed, we have begun work on bringing in psychological aspects such as goals and emotions, suitably elicited from subjects, as a basis for creating commitments and judging commitment outcomes and overall trust.

5 Acknowledgment

This work is supported by the Army Research Laboratory in its Network Sciences Collaborative Technology Alliance (NS-CTA) under Cooperative Agreement Number W911NF-09-2-0053.

REFERENCES

[1] Sibel Adalı, Fred Sisenda, and Malik Magdon-Ismail, 'Actions speak as loud as words: Predicting relationships from social behavior data', in *Proceedings of the 21st International Conference on World Wide Web*, WWW, pp. 689–698. ACM, (2012).

[2] Cristiano Castelfranchi and Rino Falcone, *Trust Theory: A Socio-Cognitive and Computational Model*, Agent Technology, John Wiley & Sons, Chichester, UK, 2010.

[3] Thomas F. Colman and Yuying Li, 'An interior trust region approach for nonlinear minimization subject to bounds', *SIAM Journal on Optimization*, **6**(2), 418–445, (1996).

[4] Andrew Fiore and Jeff Heer, 'UC Berkeley Enron email analysis', (2004).

[5] Anup K. Kalia, Hamid R. Motahari Nezhad, Claudio Bartolini, and Munindar P. Singh, 'Monitoring commitments in people-driven service engagements', in *Proceedings of the 10th IEEE International Conference on Services Computing (SCC)*, pp. 160–167, Santa Clara, California, (2013). IEEE Computer Society.

[6] Bryan Klimt and Yiming Yang, 'The Enron corpus: A new dataset for email classification research', in *Proceedings of the 15th European Conference on Machine Learning*, volume 3201 of *LNCS*, pp. 217–226, Pisa, (2004).

[7] Lauren E. Scissors, Alastair J. Gill, Kathleen Geraghty, and Darren Gergle, 'In CMC we trust: The role of similarity', in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI, pp. 527–536. ACM, (2009).

[8] Yonghong Wang and Munindar P. Singh, 'Evidence-based trust: A mathematical model geared for multiagent systems', *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, **5**(4), 14:1–14:28, (November 2010).