Capturing Reputation Features in Multiagent Systems through Emerging Patterns

(Preliminary report)

Walter M. Grandinetti[1]
wmg@cs.uns.edu.ar

Carlos I. Chesñevar[2]
cic@eps.udl.es

[1]Laboratorio de Investigación y Desarrollo en Inteligencia Artificial (LIDIA)
Departamento de Ciencias e Ingeniería de la Computación
Universidad Nacional del Sur – Alem 1253 – B8000CPB Bahía Blanca, ARGENTINA

[2]Grupo de Inteligencia Artificial – Departament d'Informàtica i Enginyeria Industrial
Universitat de Lleida – C/Jaume II, 69 – 25001 Lleida, Catalunya (SPAIN)

Abstract

Multiagent systems and online communities rely on rating systems to infer the reputation given to an individual within a particular context. The notion of reputation is essential for helping a given individual to trust in other individuals and for being himself reliable to others. Current techniques for computing individual's reputations are solely based on recent activities, facilitating a variety of possible attacks. Moreover, the amount of trust each agent has for a given context is based just on his or her reputation. In this paper we outline a new way to thwart reputation-based attacks and to detect trends in behavioral patterns based on historical data by means of knowledge discovery techniques, particularly those existing for emerging patterns.

1 Introduction and motivations

The notion of trustfulness plays an major role within online communities and multiagent systems: identities online are usually anonymous (or at most semi-anonymous), making it difficult to ensure an effective mutual cooperation. In order to minimize risks, these communities have devised different systems that intend to give a hint about the trustiness of an agent. Most of such systems are based on ratings which represent the *reputation* of a certain aspect of an agent within the community. There is a variety of rating systems, and most of them assign a *global* reputation to each agent. However, recently designed rating systems allow using *context-dependent* reputations for each agent and computing reputation not using a global value but rather considering different values of reputations for the same agent (each value taken from a different source). This is also known as *social reputation*. Although recent improvements have provided a more reliable notion of trustiness, as remarked in (Mui 2003; Sabater and Sierra 2001) many possible "attack" methodologies can still be used against rating management systems. These attacks are usually based on the fact that rating systems consider just the *most recent* reputation of an agent even when that reputation is based on multiple sources. Since, the past is quickly forgotten, a high ranked reputation agent could perform many attacks without losing his reputation level provided that he does not perform this behavior on a frequent basis. Attacks can also exploit another weak point in rating systems, namely managing context-dependent reputation efficiently. This is due to the fact that there are so many contexts to cover that it would be impossible to keep record of every possible reputation for every agent every.

In this paper we outline an approach to solve these problems using historical information, recorded from previous agent transactions. As stated before, some agents might take advantage of their high-rank reputation status to perpetrate attacks occasionally. We contend that such attacks can be identified as suspicious *patterns* of agent behavior, detectable on the basis of historical information and KDD[1]

[1]KDD stands for "Knowledge Discovery in Databases".

techniques. Clearly, there are many kinds of regularities that can be identified in the above setting. In particular, emerging patterns allow to detect as early as possible interesting trends in data. We think that such trends can be related to certain kind of attacks such as abuse of prior performance or pseudonym attack, as detailed in Section 4.

2 Trust and Reputation in Multiagent Systems and Online Communities

In everyday social activities, we rely on subjective factors (such as body language, social network, media, etc.) in order to form an opinion for a given individual. Informally, such opinion is called *reputation* and it is context-dependent in a way that the one's reputations as a computer scientist should have no influence on his or her reputation as a cook (Mui 2003). The amount of *trust* on a given individual could be directly based on his or her reputation. In the last years virtual communities have become particularly popular (as online chats rooms, electronic markets, scientific communities and virtual multiplayer game worlds). The emergence of these communities brings new ways for interaction to occur. In order to address the problem of trust within virtual communities, the following tools have been proposed:

- *Escrow Services*: They are formal institutions intended to guarantee trust (e.g., `PayPal` (Mui 2003)). However, few institutional guarantees are available except for financial institutions.
- *History Reporting*: It is a log of agent's *impressions* based on the members' interactions, this information is recorded for assessing the risk.
- *Reputation Rating System*: It is brief based on history reporting of an agent's impressions within a given virtual community.

The above tools are aimed at enhancing the level of trust among members. Escrow services, however, are not usually available and history reporting involves analyzing a great amount of data which is usually not feasible in real-time situations. For these reasons the use of *reputation rating systems* has widespread particularly within virtual communities. It has been found that one's reputation directly affect the activities and success within the community. For instance, (Dewan and Hsu 2001) reports that a seller's reputation has significant influences on his online auction prices.

Mui (2003) defines reputation as "*the perception that an agent has of another's intentions and norms*". Reputation plays a social control role, as reported in (Abdul-Rahman and Hailes 2000), influencing agents to cooperate for fear of obtain a bad reputation. A computational model of reputation is provided in (Sabater and Sierra 2001), where the reputation system is enhanced with multiple dimensions allowing modeling three kinds of reputations. In (Mui 2003) the author presents an intuitive typology of reputation as shown in Figure 1. Reputation could be classified as individual's and group's reputation. At this level, individual's reputation describes the reputation of a particular agent whereas group's reputation describes the reputation of a agent clique. For each individual, a direct and an indirect reputation could be derived. Direct reputation is based on face-to-face interaction or observations-derived. Indirect reputation is based on other trustee agents reputation. This indirect reputation is based on prior beliefs (default reasoning), group-derived which consist of a common agreement among clique members (trustworthy members) about the reputation or propagated (reputation gathered from others in the environment). In (Sabater and Sierra 2001), another approach to infer reputation status is presented using so-called *ontological reputation*, which combines different types of reputation to generate a more abstract representation, formalized as a graph structure.

Clearly, the actions of an individual agent Ag affect his or her reputation within other agents' beliefs, which propagate this belief about Ag to the community. As it is stated in (Cosmides and Tobby 1992), when facing social dilemmas (for instance, prisoners' dilemma (Axelrod 1984)), trustworthy

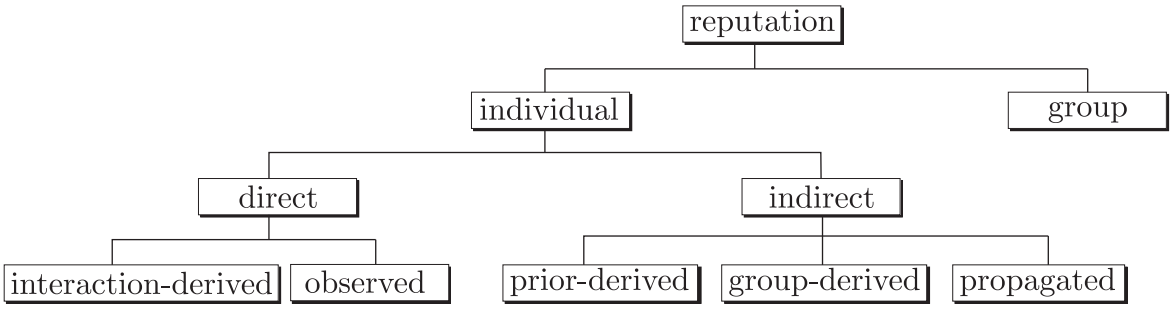

Figure 1: Reputation Typology

individuals tend to *trust* others with a reputation for being trustworthy and shun those deemed less so. An important heuristics found in human societies is the *reciprocity norm* which is also present in virtual communities. This norm states that *positive responses follow positive actions and negative responses follow negative actions*. Reciprocity norm plays an important role because –as observed in many reports– within an environment where individuals regularly perform such a norm there is an incentive to acquire a *reputation* for reciprocative actions.

As mentioned before, an agent's reputation can be seen as a collection of historical impressions about the agent. An *impression* is defined as the subjective evaluation made by an agent on a certain aspect of an outcome (Sabater and Sierra 2001). An impression $\imath$ is represented then by a tuple of the form:

$$\imath = (Agent_a, Agent_b, Outcome, \varphi, timestamp, Rating)$$

where $Agent_a$ is the judging agent and $Agent_b$ is the agent being judged, $Outcome$ reflect the particular contract and course of action of the transaction, φ is a particular variable from the $Outcome$ that is analyzed, $timestamp$ is the time when the impression was recorded and $Rating$ refers to a value in $[-1, 1]$ where -1 is completely negative and 1 is positive. For instance, consider the following dialogue on a commercial transaction, from (Sabater and Sierra 2001), between agents a and b:

$$\begin{aligned} Outcome = (&Delivery\_date =_c 10/2 \wedge Prize =_c 2000 \wedge Quality =_c A \wedge \\ &Delivery\_date = 15/2 \wedge Prize = 2000 \wedge Quality = C) \\ \imath &= (Agent_a, Agent_b, Outcome, Delivery\_date, 16:05, -0.5) \\ \imath &= (Agent_a, Agent_b, Outcome, Quality, 16:06, -0.8) \end{aligned}$$

This agreement says that the delivery date was expected to arrive $10/2$ but arrived on $15/2$, that the prize was according to the deal but the quality of the product was far below what was expected. Hence, the impressions of agent a of b about delivery variable and quality variable are negative. In the REGRET approach (Sabater and Sierra 2001) all the impressions are collected in an agent's impressions database IDB^a. The reputation for a given agent b on a given variable is computed as the weighted sum of subjective reputations *(individual dimension)*, peers reputation *(social dimension)* and ontological reputation.

3 Pattern mining

Knowledge Discovery in Databases is the non-trivial process of identifying valid, implicit, novel, potentially useful, and ultimately understandable *patterns* in data (Piatetsky-Shapiro 1991). Data mining techniques are used to find such patterns, identifiable as structures, regularities and singularities in large and growing data sets. One data mining task is indeed the identification of features

containing information which can contribute to a particular research question. There are several kind of interesting patterns to be mined. The most commonly extracted pattern is the frequent pattern which represents regularities in the data set. Other kind of patterns are infrequent, sequential, closed, maximal, discriminant, emerging, etc. In particular, we will focus our current analysis toward the usage of *emerging patterns* within a multiagent environment based on reputation rating system. However, it is interesting to see that other kind of patterns could help to understand better the domain and it could have a different application.

Emerging patterns (EPs) are associated with two data sets and are used to describe significant changes between them (Li 2001). They are especially useful to point out changes and differences between data sets, and can also capture emerging trends when applied to timestamped databases. Informally, EPs are itemsets whose support increases significantly from one dataset $\mathcal{D}_1$ to another dataset $\mathcal{D}_2$. The change in support for an itemset X is measured by a *growth rate*, defined as the ratio of X's support in $\mathcal{D}_2$ over X's support in $\mathcal{D}_1$. A typical EP example, extracted from (Li and Dong 2004), is the following: "*Lung-cancer incidence rate among smokers is* 14 *times that of none-smokers*". This example is based in two data sets, one of smokers and the other of people who do not smoke.

4 Capturing Reputation Features using Patterns

As we have discussed before the most commonly accepted tool for trust producing are reputation rating systems, which provide a reasonable trade-off between the different aspects involved in assessing trust. However, reputation rating systems fail in many situations which can be detected by considering history reporting features. As pointed out before, history reporting involves a great amount of data, and processing it is a computationally complex task. We think that techniques for detecting emerging patterns (as the ones described in section 3) can be integrated as a complement of traditional reputation rating systems. There are, classically, two kinds of attack used within a reputation rating system (Mui 2003)

- **Abuse of prior performance**: This attack consists of high-ranked rating users who take advantage of their reputation to perform some abuses without paying reputational consequences. Suppose that a user with a very good reputation begins committing fraud or defecting cooperation. Because of his high reputation, few negative ratings will not harm his reputation, cheating other agents which cannot keep track of such a behavior. Recent works try to handle this problem considering just the most recent impressions. However this trick can easily be puzzled out detecting the sliding window the reputational systems deals with. We think that user profiles can be enriched by incorporating emerging patterns that allow to identify trends on the basis of past behavior.

- **Pseudonym attack**: This attack is based on the possibility for a user to change his or her pseudonym online. That allows that the user misbehave without being detected and without paying reputational consequences. Reputational systems usually fail to discover such attacks. However, from intrusion detection research, it is known that this behavior can be detected in the same way that policemen catch a criminal following the patterns of his actions. Emerging patterns are known one of the most accurate classification methods. Based on training examples from pseudonym attacks perpetrated, the emerging patterns can accurately detect, as early as possible, this kind of behavior either from the same agent or from other agents with similar behavior.

Beside the attacks, there is another problem related current reputational rating systems. Let us suppose that a travel agent within an e-commerce community has a good rating arranging business and holiday travels all over Europe but every client that travels to Norway results very upset for the hotel accommodation. This situation does not worry the agent travel because few people travel to Norway. However, client agents looking for a trip will see this travel agent as a safe choice to go everywhere (including Norway). This is because his reputation is computed as the time weighted sum of all agent's impressions (recent impressions weighs the most). The usage of pattern mining over the set of outcomes and impressions could easily provide a pattern such $(Agent = b, Destination = Norway, Rating = Negative)$. We believe that it is important to mine not only the impressions but also the outcomes. The reason is that the outcome can be automatically generated, but the impressions rely on users who sometimes do not provide any feedbacks (as it is stated in Pollyanna effect (Mui 2003)).

Finally, the usage of patterns helps to better understand how agents behave in general and how the main quantities of interest (reciprocity, trust, and reputation) relate to each other.

5 Conclusions

Trust and reputation have emerged as an important issue in multiagent systems and virtual communities. As outlined in this paper, existing approaches using reputation rating systems have some weaknesses which can be exploited to attack them. We believe that reputation rating systems could be improved by providing a more complete description of the agent's reputation on the basis of historical data. We contend that data mining techniques as those existing for detecting emerging patterns (EPs) can be integrated with existing reputation rating systems to build a user profiles in which suspicious trends or attacks can be early detected. Part of our current research is focused on extending existing frameworks for reputation rating systems in order to include the obtention and assessment of EPs as a new feature within a formal reputation model.

References

Abdul-Rahman, A. and S. Hailes (2000). Supporting trust in virtual communities. In *HICSS*.

Axelrod, R. (1984). *The evolution of cooperation*. New York, Basic Books.

Cosmides, L. and J. Tobby (1992). Cognitive adaptations for social exchange. *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, 163–228.

Dewan, S. and V. Hsu (2001). Trust in electronic markets: Price discovery in generalist versus specialty online auctions. http://databases.si.umich.edu/reputations/bib/papers/Dewan&Hsu.doc.

Han, J. and M. Kamber (2000, August). *Data Mining: Concepts and Techniques* (3 ed.). The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann.

Li, J. (2001, January). *Mining Emerging Patterns to Construct Accurate and Efficient Classifiers*. Ph. D. thesis, The University of Melbourne.

Li, J. and G. Dong (2004). Mining border description of emerging patterns from dataset pairs. Technical report, Wright University, USA.

Mui, L. (2003). *Computational Models of Trust and Reputation: Agents, Evolutionary Games, and Social Networks*. Ph. D. thesis, Massachusetts Institute of Technology.

Piatetsky-Shapiro, G. (1991). Discovery, analysis, and presentation of strong rules. In G. Piatetsky-Shapiro and W. Frawley (Eds.), *Knowledge Discovery in Databases*. Cambridge, MA: AAAI/MIT Press.

Sabater, J. and C. Sierra (2001). REGRET: reputation in gregarious societies. In J. P. Müller, E. Andre, S. Sen, and C. Frasson (Eds.), *Proceedings of the Fifth International Conference on Autonomous Agents*, Montreal, Canada, pp. 194–195. ACM Press.