An Application of Heterogeneous Agents to Fabricate Large, Realistic Corporate Transaction Data Sets for Data Mining Tool Testing and Evaluation

Kevin B. Pratt
Computer Science Innovations, Inc. (CSI)
kpratt@csi-inc.com

Abstract

We describe methods used to specify and instantiate hundreds of heterogeneous agents and their use in a simulation of national trade and shipping. The agents, representing synthetic corporate entities, interacted to produce hundreds of thousands of trade transaction documents. The goal for the system was for the corpus of documents to evidence diverse, but realistic linkage patterns of corporate entities engaged in emergent shipping behaviors. We then used the documents to test and evaluate a data mining tool that purported to be able to detect these types of behavioral patterns. Our contributions include a design algorithm for a heterogeneous MAS that produces multi-featured outcomes, and a method for instantiating realistic heritages and preferences of agents that extends recent work in heterogeneous random utility modeling.

Keywords: Agents and complex systems, multi-agent simulation and modeling, synthetic agents: human-like, lifelike and believable qualities.

1. Introduction

A customer requested that we test and evaluate a third-party data mining tool that purported to be able to detect patterns such as those in corporate shipping behaviors. We applied heterogeneous agents to simulate the complex transactions and assemble a 15 feature output document for each transaction. To meet a qualitative standard of realism, we situated each agent in a demographically credible individual context, provided methods by which the agent could make relevant and individually appropriate decisions, allowed agents to accumulate and use experience, and enabled use of preferences and loyalties. Agents received an individual role, heritage, memory capability, personality, and decision engine.

2. Related Work

Parunak [4] reviewed the variety of practical and industrial applications of agent systems. Agents have been adapted for decision simulation in economics and industrial operations research [2]. Inclusion of heterogeneity among individuals has been an issue in utility theory and in discrete choice models [1,5]. The contemporary approach in random utility models implemented using logit (parametric) formulations is to represent heterogeneity by adjusting the value of an estimated population attribute using random selection of a normally distributed dispersion [3,5]. The approach assumes continuous distribution across the agents. It lacks behavioral explanation for the variation [3]. While latent segmentation improves heterogeneity modeling, it fails to represent complex latent decision making factors [5]. An alternative is the behavioral, exposed-process decision making that we employ for the MAS here.

3. Multi-agent system design

3.1 Decomposition of outcome

We begin with the algorithm in Figure 1 that creates and identifies agents, attributes and communication links in breadth-first fashion, starting backwards from a multi-featured outcome. The algorithm assures specification of the complete set of agents, objects and communications links. This decomposition of outcome algorithm extends backwards event planning techniques to outcome events having many inter-related features. We applied the decomposition to specify role, heritage, memory, personality and decision making processes for each agent. The system involved four agent species: manufacturer, buyer, transporter and scribe and two situated objects, contents and transport container.

3.2 Creation of realistic heterogeneity

Because this multi-agent system is too complex and large for manual attribution of feature values, we built a “factory” to instantiate the agents.

First, we obtained demographic distributions from public sources, or behavioral estimates of the distribution of attribute values. These included demographics of larger cities, annual contents shipments in dollars, and two distinct distributions of corporate gross sales. We used a behavioral estimate of the distribution of corporate purchaser loyalty and domain expert estimates of the distributions of
Initialize: create null set of agent_species, and for each, a null set of attributes and a null set of communication_links; create null queue of features.

for each feature in outcome, enqueue feature.

while queue not empty, dequeue feature

identify who makes the decision that puts a value in the feature. create, or identify an existing agent_species to make the decision.

create communication_link from agent_species to feature.

identify the meta-type of the decision. create or identify an agent role attribute.

identify the information that is relevant to this decision meta-type

for internally accessible information, create, or identify an agent attribute.

for externally accessible information, define that information to be a feature and enqueue the feature.

Figure 1. Algorithm for decomposition of outcome.

Define $S = n$-dimensional feature space, each dimension $i$ having domain $[0, \ldots, b_i, \ldots, m_i]$ where $b_i, m_i \in \Gamma$. Assume there exists a population $A$ of elements in $S$, where more than one element can exist in a unit of $S$. There exists a probability of occurrence $p(A)$ everywhere in $S$.

Further, there exists in $S$ a set of points $R$, where $R$ is defined as follows:

$R = \{ r_i = (r_{i1}, \ldots, r_{ij}, \ldots, r_{in}) | \text{where } r_{ij} = \text{random}[0, m_j] \text{ for all } j, \text{ and } p(A \text{ at } r_i) > t \text{ where } t = \text{random}[0,1] \}$.

$R$ has the property that as the cardinality of $R$ increases, $p(R) \rightarrow p(A)$ everywhere in $S$.

Assertion: $R$ is an arbitrarily accurate stochastic estimation of the heterogeneity of $A$.

Figure 2. Discrete stochastic heterogeneity model.

shipping container ownership, and container types. Next, the factory method instantiated agents with the attribute values in three steps. It assigned any necessary predicate demographic values to the agent’s heritage from a demographic distribution. It selected the number of attribute values required for each multi-valued attribute. Finally, it filled those values conditional on the individual agent’s demographic heritage.

We formalize this technique as the discrete stochastic heterogeneity model in Figure 2. As the number of agents having a demographic based attribute increases across the system, the match to the real-life demographics will improve. We can apply any demographic distribution(s) of any shape, without limiting assumptions such as normality or continuity. Escaping the requirement of continuity is critical when modeling un-ordered, categorical data types, where joint probabilities may be discontinuous everywhere and latent segments may be difficult to define.

4. Simulation and results

An external trigger motivated a buyer to buy, based on the distribution of buyer sales, thus large buyers bought and shipped much more often than small buyers. The necessary decisions by various agents then cascade until scribe can write the completed shipping document.

The buyers and manufacturers evolve their decisions as they gain experience from interaction with each other. These agents use memory of experiences of past transactions and degrees of loyalty. Patterns emerged. Based on a buyer’s heritage, it bought more often for its warehouses in large cities, and bought more of those contents it needed. The buyer often bought from manufacturers who it had dealt with before expressing loyalty. On the other hand, a manufacturer sometimes refused sales to buyers with whom it had bad experiences. We also used the factory to add rogue agents to the system.

5. Conclusion

We designed and applied a large MAS to a significant and complex customer problem. Heterogeneity among heritage and preferences of individual agents was modeled realistically using a novel method to directly apply demographic distributions.

References:


