WCCI 2010 IEEE World Congress on Computational Intelligence
July, 18-23, 2010 - CCIB, Barcelona, Spain IJCNN

Building the Knowledge Base of a Buyer Agent Using Reinforcement Learning Techniques

George Boulougaris, Kostas Kolomvatsos, and Stathes Hadjiefthymiades

***Abstract*—Electronic markets are places where entities not known in advance can negotiate and agree upon the exchange of products. Intelligent agents can be proved very advantageous when representing entities in markets. Mostly, such entities are based on reputation models in order to conclude a transaction. However, reputation is not the only parameter that they could be based on. In this work, we deal with the problem of how and on which entity a buyer should be rely upon in order to conclude a transaction. Reinforcement learning techniques are used for these purposes. More specifically, the Q-learning algorithm is used for the calculation of the reward that the buyer will take for every action in the market environment. Actions represent the selection of specific entities for the negotiation of products. The most important is that the reward values are calculated based on a number of parameters such as the price, the delivery time, etc. The result is a more efficient model that is not based only on the reputation of each entity. Finally, we extend the Q-learning algorithm and propose a methodology for the dynamic Q-table creation which results reduced time for its construction and respectively limited time for the purchase action. Simulations show that this model indicates a significant time reduction in the purchase process in conjunction with the best solution according to the characteristics of products.**

I. Introduction

Nowadays, users are in front of a huge amount of information sources as well as product resources. Due to the numerous resources, the discovery and the purchase of the appropriate products becomes a task that is out of the human capabilities. The main reason is that users should spend a lot of time and effort for searching among of millions of resources to find products that fully satisfies their needs. For this, an automatic tool for the discovery of product information is necessary. Such tool seems to be Intelligent Agents. Agents are components capable of acting autonomously in order to achieve goals defined by their owners. Their intelligence mostly refers to their capability to learn the preferences of their owners, thus, increasing their performance. Hence, agents can undertake the responsibility of finding product information in the Web with the minimum users' intervention. Furthermore, agents can be very advantageous when representing entities acting in the Web. For example, they can undertake the responsibility of negotiating the purchase of a product when they discover it to match their owners needs.

Agents can participate in places where they can negotiate and agree upon the exchange of products. Such places are Electronic Markets (EMs). In EMs, usually, there are three main groups of participants: the buyers, the sellers, and the mediators. Buyers are entities that search to purchase the product that best matches their needs. Buyers want to buy products at a profitable price which is the lowest possible one. At the opposite side, sellers have a number of products in their property and they want to sell them in the most profitable price which is the highest possible. Mediators are mainly used for administration purposes. For example, they can be used for payments facilitation, trust calculation, etc. The combination of EMs with Intelligent Agents can be very advantageous for discovering and acquiring products. Agents can represent participants in EMs facilitating the automatic process about the purchase of goods. In this work, we focus on the interaction between buyers and entities negotiating products and especially in their selection process. Such entities can be brokers or sellers. Brokers are mediators that undertake the responsibility of finding and returning the desired products to the buyers. We study a technique based on which buyers can identify the most appropriate entity in order to rely on it concerning the purchase of a product. In the rest of the paper, when we refer in 'selling entities' we refer in entities having the capability of negotiating over a number of products for specific returns (sellers or brokers).

Usually, buyers rely upon the reputation of a specific entity in order to conclude a transaction. However, buyers should take into consideration a number of other parameters such as the price, the delivery time, the product relevancy with its goals, etc. The ideal product for a buyer is the product that can be sold at the minimum price, can be delivered in the smallest time and it is highly relevant to its goals. In this research effort, we take into consideration all these parameters in the buyer decision process. Our model enables reinforcement learning techniques [1], which provide a general framework for sequential decision making problems and they are proved efficient for many important applications. Reinforcement learning deals with how an agent should take actions, at every state that it can be, in order to have the maximum long term reward. These algorithms discern states of the world and the agent can choose the best policy according to the state that is. More specifically, Q-learning methodology works by learning the reward that the agent will take for a given action in a given state of the world and following a specific policy. The discussed algorithm works without the need of modeling the world.

Specific tables used in the Q-learning methodology will provide a 'knowledge base' in the buyer's side, which will

Authors are with the Pervasive Computing Research Group, Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Panepistimiopolis, Ilissia, Athens, phone +302107275127; e-mails:{grad0705, kostasks, shadj}@di.uoa.gr. Corresponding author is the second author.

lead to the best action at every state. We show that this model requires the minimum number of steps for the conclusion of a transaction. The discussed methodology is important, because it provides a simple and clean language to state the specific problem and to lead to the best solution. The buyer observes its state and the other buyers' actions and decides, according to the long-term reward maximization, the specific action. Moreover, they behave optimally even more in noisy and very dynamic environments.

The rest of this paper is organized as follows: Section II reports prior work while in Section III we give the necessary description of EMs and Intelligent Agents. Section IV is devoted to the description of our model which involves the usage of reinforcement learning techniques in the interaction between buyers and selling entities. In Section V, we discuss our results indicating that we have a large reduction in the time required for each purchase process. Finally, in Section VI, we conclude the paper by presenting some future extensions.

II. Related Work

In literature, one can find some very interesting approaches concerning EMs and their usage. In [2], authors describe an EM where intelligent agents act as buyers, sellers or intermediaries. An intermediary agent coordinates information related to buyers, sellers and what is being offered or demanded. The intermediary agent also handles other important issues like interaction security, fraud and payment handling.

MAGMA is presented in [3] and it concerns of an agent-based market architecture where agents can buy or sell products. Two modes are available for negotiation: manual and automatic. Administration issues are defined for product storage/manipulation and economic mechanisms for payments. Moreover, an advertisement server is available and a relay server facilitates the communication between trader agents.

In [4], authors examine the role of middle entities in information markets. Information markets are EMs where entities negotiate over the purchase of information products. They simulate an electronic market where agents play basic roles, e.g. information producers, information suppliers and information mediators. These mediators are called InfoCenters. Their results indicate that middle entities can enhance the market's efficiency, however, they didn't affect market prices. Their main advantages concern facilities provided to other market participants.

In [5], authors present a survey on agent-mediated electronic commerce systems. They describe the increasing role of agents as mediators in electronic commerce. Their work explores these roles, their supporting technologies as well as how they relate to electronic commerce. The main points that they focus on are business to business, business to consumers and consumers to consumers transactions. They study agents in the context of a Consumer Buying Behavior (CBB). The CBB model augments traditional marketing techniques with agent research efforts to accommodate electronic commerce.

Another extensive survey on agent-mediated e-commerce is presented in [6]. Authors focus on B2C and B2B aspects of the electronic commerce. They extend previous efforts on the field by presenting a more integrated and coherent view on the discussed issues. The roles of agents as buyers, brokers or negotiating entities are described and studied. The CBB model is followed as well as the B2B Transaction (BBT) model. Moreover, authors extend the traditional CBB model in order to cover more B2C behaviors.

A reinforcement learning model in electronic markets is presented in [7]. It concerns a continuous learning mechanism that agents adopt when negotiating for resources. Authors propose a strategy which quickly converges leading to Nash equilibrium when agents face other adaptive opponents. In their simulations, they examine the case where buyers try to buy services from sellers using a first-price sealed bid auction.

Authors in [8] deal with software agents that utilize reinforcement learning algorithms to make their decisions in a marketplace. The Q-learning algorithm is used either by one or both of the competing agents trying to decide their best policy concerning the proposed prices. In the first case, the agent using the learning mechanism yields greater expected profit than the other. When both of agents use the learning mechanism there is no proof for convergence, however, the proposed architecture yields good performance.

A personalized agent system based on the Q-learning technique is described in [9]. This system is used for travel recommendations that match the users interests. Two learning approaches are presented in the discussed effort: In the first, the personalization learner learns from all users in one cluster to find cluster interests of travel information by using data related to ages and genders. In the second, the learner studies the user profile, the user behavior as well as trip features in order to provide the unique interest for each user.

Reinforcement learning techniques could be used for determining dynamic prices in a market scenario as shown in [10]. In this research effort, a single-seller and a two-seller scenario are examined. Sellers utilize the Q-learning algorithm in order to be able to define dynamic prices when acting in the market. More specifically, in the two-seller case, authors model the discussed problem as a Markovian game providing specific formalisms. They solve the problem by using actor-critic algorithms through simulation. Finally, the illustration of their approach is done through examples of typical retail markets.

In [11], authors describe strategies followed by buyers and sellers in order to conclude efficient transactions. Especially, buyers strategies involve the selection of the seller that maximizes their profits. They study strategic equilibria for buyers to be implemented in an automatic way.

In [12], a web based auction scenario is presented. Authors deal with the dynamic decrement of prices based on a mechanism that finds the maximum total expected revenue. This mechanism is modeled as a single reinforcement learning agent acting in an uncertain auction environment. A finite horizon Markov process is defined under the assumption of independent bidder valuations and is solved using the Q-learning algorithm. The aim is to define the arrival patterns of bidder and their price – demand curves.

III. Electronic Markets, Agents And Learning

An EM can be seen as a virtual place where entities not known in advance can interact and negotiate over the purchase of products. Products can be electronics, cloths, or even more information products. Information goods could be images, videos, music, software code and electronic articles. In such places, we can discern the following types of members: Buyers, Sellers and Mediators. Buyers trying to buy products have a specific valuation about each of them and they are not willing to pay more. Sellers have a number of products in their property and try to sell them in the most profitable price. Sellers have a specific production cost and they are not willing to sell products in prices below this cost. Mediators are mainly used for administration purposes. These entities can facilitate buyers and sellers in order to interact in this market. Mainly, mediators can be discerned as brokers or matchmakers. Brokers can undertake the responsibility of finding and returning the appropriate product related to the buyer's needs while matchmakers result a seller's address based on buyers requests. It should be noted that matchmakers cannot sell products as brokers do. In this paper, we focus on the interaction between buyer and entities selling products. These selling entities can be sellers or brokers. We focus on the selling entity selection process and through reinforcement learning techniques we provide an efficient way for the purchase action. The description of the communication protocol between buyers and selling entities is out of the scope of the current paper.

Most of the research efforts, dealing with the problem of the selection of choosing a seller or a middle entity, focus on the usage of reputation. However, the reputation is not the only reason for a buyer to decide to negotiate with a specific entity. Let us think the case where two or more sellers or middle entities have the same reputation level. In such cases, the buyer should decide based on a number of parameters such as the price, the delivery time, etc. However, this implies a computational effort in the buyer's side. Every time there is the need for a purchase the buyer should interact with the entities that trust and accordingly to decide the entity with which it will negotiate or conclude the transaction.

In the discussed scenario, learning techniques can help buyers to identify the entity they can rely on, in order to buy products. Reinforcement learning [1] is a sub-field of machine learning. It deals with the behavior of an agent that tries to take some actions in an environment and being at some state. Through actions the agent tries to maximize its long-term reward. Hence, it tries to find a policy that maps states to actions. Q-learning is a reinforcement learning technique. The main advantage of this algorithm is its capability to define the expected utility without the need of modeling the environment. In its simplest form, the algorithm uses tables to store data related to the rewards that the agent will gain following a policy.

Reinforcement learning differs from supervised learning as correct inputs/output pairs are never presented nor sub-optimal actions explicitly corrected. In reinforcement learning there is a focus on online performance, which involves balancing between exploration and exploitation. The basic reinforcement learning problem as applied to Markov Decision Processes (MDPs) consists of: a) A set of environment states S, b) A set of actions A, and c) A set of scalar rewards defined in $\Re$.

At each time step the agent perceives at being in state $s_t \in S$ and the set of possible actions $A(s_t)$. It chooses an action $\alpha \in A(s_t)$ and receives from the environment a new state s_{t+1} and a reward r_t. It should be noted that the discussed scenario involves a very dynamic environment because the agent should interact with entities that change their characteristics continuously (prices, delivery time, etc). Based on these interactions the reinforcement learning agent should develop a policy $\pi : S \rightarrow A$ which maximizes the quantity:

$$R = r_1 + r_2 + \ldots + r_n \quad (1)$$

for MDPs that have terminal state, or the quantity:

$$R = \sum_t \gamma^t \cdot r_t \quad (2)$$

for MDPs without terminal states. We have:

$$0 \le \gamma \le 1 \quad (3)$$

The factor γ is the future reward discount factor. Hence, the buyer agent B at every time step perceives its environment state and decides to perform a specific action according to this state and the feedback from previous actions. The buyer acts towards to a specific goal which involves the purchase of the appropriate product. The appropriate product has specific characteristics involving: a) the price, b) the delivery time, c) the relevance with the buyer's goals, and, d) the minimum number of steps required for the transaction (purchase time).

IV. Reinforcement Learning In A Virtual Market Scenario

A. Scenario

Trying to use the advantages provided by reinforcement learning, we develop a virtual marketplace where potential buyers utilize Q-learning in order to learn on which entity they can rely on for their purchases. We take into account the assumption that the buyer interacts only with entities having a high reputation degree. In our model, the following entities participate: the entity selling products (broker or seller) and the buyer.

Each selling entity is an intelligent agent having the capability to negotiate the purchase and the delivery of a product. There is a number of entities acting in the marketplace. They can negotiate a maximum number of products. For each selling entity, information related to every product is:

- The product id. It is necessary in order to be uniquely identifiable.
- The time for which the product is valid. After this time limit the product is considered obsolete and has no valuation. For example, a stock price is valid only for a limited period of time.
- The product price. We consider that this price is final and it contains the entity's fee (in the case of a broker).
- The time in which the product will be available to the buyer. For example, a video file may have such duration that a nontrivial amount of time is necessary for the delivery.
- The product relevance to the buyer's goals. This is calculated either by the buyer or by a marketplace entity used for such purposes. The calculation is based on an algorithm imposed by the buyer in order to have an objective view. For example, the relevancy factor could be a real number in the range [0..1] indicating the product relevance to the buyer's goals. For example, when a buyer searches for operating systems software product, then a product '*SomeDistribution Linux 9.2*' has high relevancy score close to 1 and the opposite stands for the product '*SomeTextProcessor 4.5*'. The description of such algorithm is out of the scope of the current work.

It should be noted that the number of products differs among entities.

The buyer represents a user and tries to buy some products relying on a selling entity. Actually, a buyer is consisted of two parts. The first part is responsible for the creation of the Q-table according to entities and products characteristics. We consider that information related to entities can be found relying on specific marketplace intermediate entities. Moreover, we consider that the time required for the communication between the buyer and this middle entity is negligible. When it is necessary (i.e. a product is not available), the buyer updates the appropriate elements of the table (actually it updates the appropriate table – see next Section). The second part of the buyer is responsible for the completion of a purchase action. When there is the need for a purchase, the buyer uses the Q-table in order to choose the appropriate entity and accordingly concludes the transaction.

B. *Q-Learning Table Creation*

The table creation process is held in an initial step, when the buyer starts to act in the market. The buyer has a Q-table for each product. The table has two dimensions. Rows represent the states of the world (with which entity interacts) and the M+1 columns represent actions that the agent could take, where M is the selling entities number. This means that when the buyer is at a specific entity, it decides what action will take according to the values of the specific row of the table. This decision could be the purchase of the product (action M+1) or the transition to the next entity that corresponds to the specific action and consists of the best choice at the current state. In other words, the rejection of the purchase action and the selection of another entity corresponds to a '*not buy from this entity*' action. Hence, the number of all the Q-tables elements is given by the following equation:

$$\mathrm{Elements} = \mathrm{P} \cdot (\mathrm{M}+1)^2 \tag{4}$$

where P is the number of products.

Figure 1 shows an example snapshot of this complex Q-table.

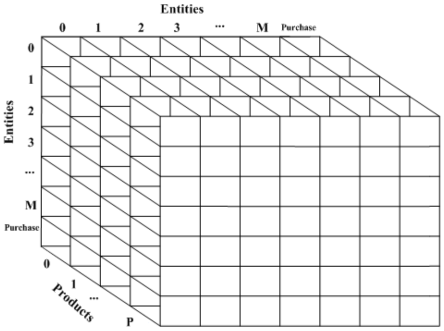

Fig. 1. Q-table general form.

Information that the buyer takes into consideration in order to build the Q-table is: a) Product Relevancy Factor (*R*), b) Product Price (*Pr*), c) Response Time (*RT*), and, d) Number of Transitions (*NT*). The parameter *R* indicates how much relevant is the product to the buyer's goals while *Pr* is the return that the entity asks for the specific product. It is obvious that the most possible case is that sellers sell their products in smaller prices than brokers and this way it is very possible to be beneficial through the Q-table creation algorithm because this is going to be highlighted from the Q-table values. The reason is that the algorithm takes into consideration all the above mentioned parameters using specific rewards for each of them. *RT* indicates the time in which the purchase will be completed and the product will be delivered to the buyer, and *NT* shows how many transitions the buyer will need in order to conclude a decision (selection of an entity). For every parameter, we have defined a methodology that results its final value (see Section V). For example, the smaller the number of transitions is the greater the reward becomes. This is because the buyer wants to conclude its interaction at the smallest number of steps. Also, if the entity's proposed price is smaller than the half of the buyer's valuation the reward that the buyer gains is greater than in cases where the opposite stands. Similar rationale stands for the rest of parameters.

Q-table is created based on the following equation [1]:

$$\mathrm{Q}(\mathrm{s}_t, \mathrm{a}_t) \leftarrow \mathrm{Q}(\mathrm{s}_t, \mathrm{a}_t) + \mathrm{l} \cdot \left\lfloor \mathrm{r} + \gamma \cdot \max_{\mathrm{a}' \in \mathrm{A}} \mathrm{Q}(\mathrm{s}', \mathrm{a}') - \mathrm{Q}(\mathrm{s}, \mathrm{a}) \right\rfloor \tag{5}$$

where l is the learning rate, r is the reward, γ is the future reward discount factor, $Q()$ gives values taken from the Q-table, and s_t and a_t are the state and a specific action at time t respectively. States and actions are represented by an entity in the market. Every selling entity represents a state in the table and actions are the transitions to other M-1 entities or the purchase of the product. In our model, we define a reward decrement for states leading to entities not having the specific product. The learning rate indicates when the learning process has reached to the final point. The learning rate is calculated based on the number of episodes and decreases over time.

Concerning the calculation of the Q-values, it should be noted that the Q-table rewards are decreased by a specific value (we have used a 5% decrement) when deal with entities not having the specific product. Moreover, the reward r is based on three partial rewards: a) the reward for the relevancy factor, b) the reward for the product price, and, c) the reward for the response time. It should be noted that:

- The greater the relevancy factor is the greater the reward becomes.
- The smaller the price is the greater the reward becomes.
- The smaller the response time is the greater the reward becomes.

As we mentioned, the Q-table indicates the most appropriate selection of a specific action in a specific state. However, in dynamic environments such as an EM, buyers should act under different conditions after time to time. For example, a product could not be available in a specific selling entity or an entity could not be available for transactions. In such cases, the buyer should act immediately and update the Q-tables. In this point, we consider that when such an alteration in market state happens all buyers are informed immediately by the appropriate market intermediate entity. We consider that the required time for buyers to be informed is negligible.

In this point, we extend the Q-learning methodology and when a product is imported or revoked, the buyer updates only the elements that refer to the specific product and the specific entity. This way, the buyer saves time. Moreover, every time a new product enters to the market the buyer creates a new table for this product. In our scenario we deal with the following cases:

- A new product is available in the market.
- An entity can negotiate for a new product that it is available to other entities.
- An entity cannot negotiate for a specific product any more.
- A product is not available by any entity any more.
- An entity is not available for any negotiations.

Furthermore, a specific episodes number is used in the training phase. The training phase aims to the productive creation of the Q-tables. We need a large enough episodes number in order to have efficient Q-tables. However, we adapt the Q-tables creation algorithm in such a dynamic environment by using dynamically calculated episodes number every time the training phase starts. If we choose to have a constant episodes number, probably this will be not effective especially in cases where we have to update the Q-tables many times. Updating the Q-tables many times would be the normal case in dynamic markets where large number of entities acting in them. As mentioned, when there is the need for the Q-tables update, the buyer only deals with the table and rows related to the specific product. Hence, a dynamic calculation methodology would be more appropriate for this scenario. In this paper, we propose a method for the dynamic episodes calculation according to the characteristics of the market. The equation used is:

$$NE = c \cdot (M + 1)^2 \cdot P \qquad (6)$$

where M is the number of selling entities, P is the maximum number of products for each entity and c is a constant parameter. The value of c is taken by simulations. The above equation indicates that the number of episodes could be a small number in cases where there are a few entities and each of them has a small number of products. The time that we save in such cases is very important because it leads to the overall reduction in the table creation time and respectively in the general purchase time.

C. Buyer purchase behavior

In the literature, we can find a lot of research efforts that study the behavior of a buyer in a market scenario. However, the majority of them involve a scenario where the buyer should interact with a number of selling entities either in a sequential order or in parallel. In such cases, the buyer should interact first with the entities and accordingly decide the purchase action. Furthermore, when a buyer interact with an entity, it is not feasible to know the characteristics of the rest of the entities (price, delivery time, etc) in order to be able to choose the most profitable purchase action. The importance of our proposed methodology is that the buyer using the Q-learning algorithm is able to incorporate the knowledge about all the entities and their characteristics for every product in the Q-table. Hence, it is able to decide the best solution at every state and to adapt immediately to alterations in market's characteristics. Moreover, it saves time because it is not forced to negotiate with all the selling entities and accordingly to decide the entity that can rely upon.

The first step of the buyer purchase behavior is to collect the necessary information and create the Q-table. After creating the Q-table, the buyer is able to buy products. At first, it randomly selects an entity for interaction for the specific product and tries to conclude a purchase. This entity represents the initial state of the buyer. If this is the i-th entity, the buyer looks at the i-th row of the appropriate Q-table. Based on the values retrieved by the i-th row, it is able to choose the maximum value and, thus, to choose the action that should take. This action could be the purchase of the product from the current entity or the transition to another one. If the selected entity does not have the requested product or the buyer learns from the market entities that a product is revoked, it chooses its next best transition and accordingly updates its table. If the best indicated action is the return to a previous visited entity that had declared inability to deliver the product, the purchase is not feasible. In such cases, it is not profitable for the buyer to purchase the product due to the entity's characteristics (price, relevance, response time) that result the current value of the Q-table.

V. Results And Discussion

In our model, the total purchase time is equal to the summary of tables creation/update time and the time required for the purchase decision. Hence, the following stands:

$$T_p = \sum_{i=1}^{T} T_{ci} + P \cdot \sum_{j=1}^{K} T_{tj} \qquad (7)$$

where T_p is the total purchase time for a specific product, T_{ci} is the time required for the i-th table creation/update time, T_{tj} is the time required for the j-th transition in tables rows, T is the number of tables, K is the number of the transitions for the purchase of the specific product and P the number of products. In (7) the time required by each transition in tables rows depends on the steps required for each transaction. In the Q-learning approach we calculate the expected number of steps that is:

$$E_Q(\text{steps}) = (M-1) \cdot \frac{1}{M} \cdot 2 + \frac{1}{M} = 1 + \frac{M-1}{M} \qquad (8)$$

where M is the number of selling entities. The greater the M is the closer to 2 the number of steps become. However, using the Q-learning approach, the buyer, in the worst case, needs approximately 2 steps, which means that the buyer concludes every transaction for a specific product in at most 2 steps.

Without using the Q-learning approach, when buyers need to buy a product, they should interact in M steps with all the entities and at an additional step should decide the purchase action. Hence, the required steps in such cases are:

$$E_{\tilde{Q}}(\text{steps}) = M + 1 \qquad (9)$$

In these cases, we consider that each buyer is not based on its history for its purchases. This is very important, because at every time, there is a possibility that another entity can sell a specific product in a more profitable price, or delivery time, etc. Hence, the buyer should not probably rely on his history and buy products from the same entity. Moreover, if the buyer decides to ask all the selling entities for a product and accordingly decide the purchase action, the total required time for the purchase of P products is given by:

$$T_L = (M+1) \cdot P \cdot t_m \qquad (10)$$

where M is the number of entities and t_m is the interaction time with each entity. The time t_m represents the time that the buyer needs to be informed by the entity for the information related with the products that the buyer wants.

We have conducted a set of simulations for various combinations of values for basic parameters of our scenario trying to simulate a very dynamic market where changes happen either in the number of the selling entities or in the number of products. These simulations concern dynamically calculated episodes number as described in Section IV.

Moreover, we consider that at randomly selected points there are changes in the number of products that entities negotiate. The most important is that we consider cases where entities negotiate products that they do not previously deal with. This happens either when these products are available to other entities or are negotiated for a first time in the market. In our experiments, we consider a probability of 2% for the case where a new product is available in an entity and of 5% for the case where a product is totally new in the market. Also, a probability of 5% is used for the case where a product is not available in a specific entity. On the other hand, we consider in cases where a product is no longer available in an entity and totally in the market forcing the buyer to adapt his behavior in the new situation. Furthermore, in our experiments we take into consideration cases where an entity starts to negotiate for a first time in the market. The probability used for this case is 2%. Finally, we examine the case where an entity is not available for negotiations in the market under the probability of 1%. From the above, it is understood that we consider a dynamic market where entities and products enter or are revoked by the market. Our aim is to show the efficiency of the reinforcement learning techniques in such scenarios.

In our simulations, we examine purchases of 400 products and study the required time. We compare this time with the time required for purchases in the case where the buyer interacts with all the entities before it decides the purchase. In each experiment, we define a maximum number of entities in the market as well as a maximum number of products for each entity. Moreover, we define basic parameters important for the creation of the Q-table as follows (see equation 5):

- the parameter γ is defined to be equal to 0.8.
- the reward r is defined as the summary of the reward of the required movements for the purchase conclusion, the reward of the product price and the reward of the response time. Equations (11) to (14) describe the reward calculation for each parameter.

$$r = r_{\text{steps}} + r_{\text{price}} + r_{\text{time}} \qquad (11)$$

$$r_{\text{steps}} = \frac{c}{k} \qquad (12)$$

$$r_{\text{price}} = \begin{cases} \dfrac{T_{pr} - Pr}{T_{pr}} \cdot w_p, & \text{if } Pr \le T_{pr} \\ 0, & \text{if } Pr > T_{pr} \end{cases} \qquad (13)$$

$$r_{\text{time}} = \begin{cases} \frac{T_t - RT}{T_t} \cdot w_t, \text{if } RT \leq T_t \\ 0, \qquad\qquad \text{if } RT > T_t \end{cases} \tag{14}$$

In the above equations, c is a constant value, k is the number of movements for the purchase conclusion, Pr is the product price, T_{pr} is a price threshold (i.e. equal to the half of the buyer valuation), RT is the response time, T_t is a time threshold (i.e. equal to the half of the buyer deadline) and w_p and w_t are the weights to the final reward of the price and response time respectively.

- the learning rate l is initially defined equal to 0.8. However, we choose to reduce this value as the number of episodes increases. When the episodes number is very large then the learning rate is a very small number.

First of all, we examine the tables creation time. Tables I, II, and III show this time for different entities and products numbers. We depict two time indicators. The time required for the creation of the first table and the time required for the update process of the tables where there is such need. As we can see, the update process time is less than the first table creation time, because in our model we only deal with the table and rows concerning the specific product.

Table I. Tables creation time for different entities number (1).

Entities Number (5 Products each)	First Table creation time (ms)	Average tables creation time (except first table) (ms)
4	15	0
15	125	17.86
50	1685	402.73
100	16520	3546.44
200	208088	41846.64

In Tables IV and V, we depict the number of moves required for the purchase action. We compare the total number of moves for the purchase of 400 products with the number of moves required when the buyer should interact sequentially with all the entities in the market in order to choose the best solution. We can see that in average we need 1.72 moves for every purchase when using our model in contrast to a very large number of moves in the second case. The time reduction using Q-learning is very large especially when a large number of entities act in the marketplace.

Some interesting observations can be derived by Figures 2, 3 and 4. In Figures 2 and 3, we can see that the average price and the average response time are decreasing as the entities number increases. This is because the Q-learning approach takes into consideration all the necessary parameters in order to lead to the most appropriate entity for a specific product. The most appropriate entity is that having the smallest possible price, the smallest possible response time, the highest possible product relevance and is can be purchased at the smallest possible moves. Concerning, the average relevancy and the average number of moves, we do not observe any important variations for any entities or products number. This is very important because this means that the average relevance for the purchased products is at high values irrelatively to the selling entities number. The buyer at every scenario is able to choose the best possible products based on the Q-learning tables at the minimum number of steps. Furthermore, in Figure 4, we can discern that there are not any important variations in the average price or the average response time. The number of products does not affect the characteristics of each purchased product.

Table II. Tables creation time for different entities number (2).

Entities Number (40 Products each)	First Table creation time (ms)	Average tables creation time (except first table) (ms)
6	156	15.50
15	561	114.33
30	3510	453.73
60	32667	2386.50
100	191303	8254.22

Table III. Tables creation time for different products number.

Products Number (15 Entities)	First Table creation time (ms)	Average tables creation time (except first table) (ms)
5	125	17.86
40	561	114.33
80	1029	210.60
150	1731	374.57
500	6155	1319.14
1000	14917	2940.86
5000	193644	14914.00

Table IV. Transitions to purchase for different entities number (1).

Entities Number (5 Products each)	Total moves for 400 Products	Total moves for 400 Products (without using Q-learning)	Moves reduction using Q-learning
4	653	2000	-67.35%
15	716	6400	-88.81%
50	714	20400	-96.50%
100	732	40400	-98.19%
200	768	80400	-99.04%

Table V. Transitions to purchase for different entities number (2).

Entities Number (40 Products each)	Total moves for 400 Products	Total moves for 400 Products (without using Q-learning)	Moves reduction using Q-learning
6	718	2800	-74.36%
15	705	6400	-88.98%
30	703	12400	-94.33%
60	693	24400	-97.16%
100	712	40400	-98.24%

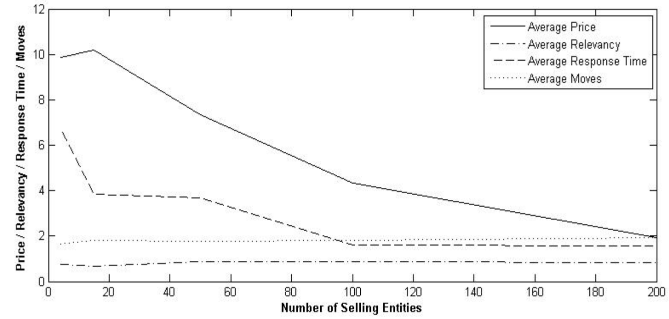

Fig. 2. Graphical representation of various parameters for different number of entities (maximum 5 products each).

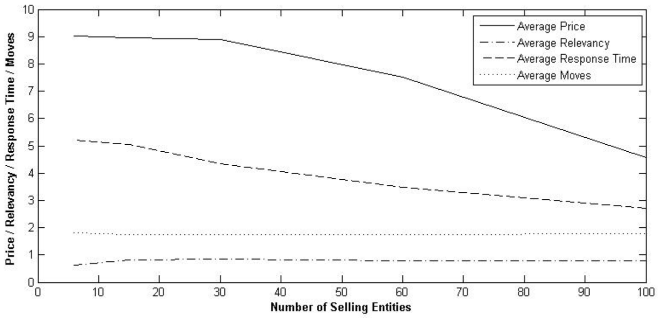

Fig. 3. Graphical representation of various parameters for different number of entities (maximum 40 products each).

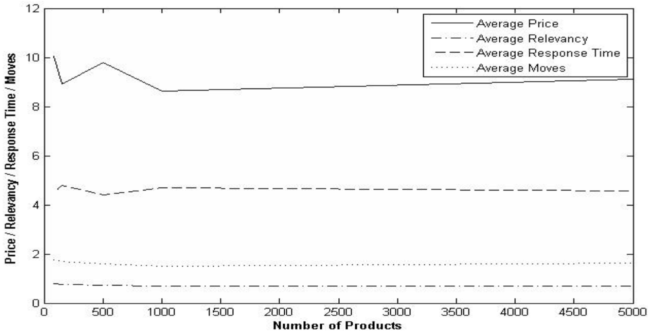

Fig. 4. Graphical representation of various parameters for different number of products (maximum 15 entities in the market).

VI. Conclusion

In this paper, we present a model for an Electronic Marketplace. In this market, there are three main types of entities: The buyers, the sellers and the mediators. We examine the case where buyers representing users interact with mediators acting as brokers or with sellers negotiating for a number of products. Buyers are able to use the Q-learning technique in order to model selling entities and through this way to be able to decide at every state which is the best action to take. Each state is an entity from which the buyer asks a product. If the purchase is not possible in a specific entity, the buyer based on the Q-table chooses its best action to the current situation.

We also present our results indicating a reduction in the time required for the completion of each interaction due to the limited number of steps required for the transaction. This time, in the Q-learning approach, mainly depends on the time required by the table creation, thus, we propose a method for the dynamic calculation of the episodes number every time that the Q-tables are created or updated. Concerning, the steps required for the purchase decision of the buyer, this stands at low levels and in the worst case is equal to 2.

References

[1] R. S. Sutton, & A. G. Barto. Reinforcement Learning: An Introduction. MIT Press, 1998.

[2] G. P. Barbosa, & Q. B. Silva. An Electronic Marketplace Architecture Based on Technology of Intelligent Agents and Knowledge. Lecture Notes in Computer Science, "E-Commerce Agents, Marketplace Solutions, Security Issues, and Supply and Demand", 2001, pp 39-60.

[3] M. Tsvetovatyy, M. Gini, M. Mobasher & Z. Wieckowski. MAGMA: An Agent-Based Virtual Market for Electronic Commerce. Applied Artificial Intelligence, vol. 11(6), 1997, pp 501-523.

[4] I. Yarom, J. S. Rosenschein, & C. V. Goldman. The Role of Middle-Agents in Electronic Commerce. IEEE Intelligent Systems, vol. 18, no. 6, pp. 15-21, 2003.

[5] R. H., Guttman, A. G., Moukas, & P., Maes, 1998, 'Agent-Mediated Electronic Commerce: A Survey', The Knowledge Engineering Review, vol. 13(2), pp. 147-159.

[6] M., He, N. R., Jennings, and H. F., Leung, 2003, 'On Agent-Mediated Electronic Commerce', IEEE Transactions on Knowledge and Data Engineering, vol. 15(4), pp. 985-1003.

[7] E. Oliveira, J. M. Fonseca, & N. R. Jennings. Learning to be Competitive in the Market. In Proc. of the AAAI Workshop on Negotiation: Settling Conflicts and Identifying Opportunities, Orlando, Florida, USA, 1999.

[8] G. Tesauro. Pricing in Agent Economies Using Neural Networks and Multi-Agent Q-Learning. In Proc. of the Workshop on Learning About, From and With other Agents (IJCAI '99), 1999.

[9] A., Srivihok, and P. Sukonmanee, 2005, 'E-Commerce Intelligent Agent: Personalization Travel Support Agent Using Q-Learning', in Proc. of the 7th International Conference on Electronic Commerce, Xi'an, China, pp. 287-292.

[10] C. V. L., Raju, Y., Narahari, & K., Ravikumar, 2003, 'Reinforcement Leanring Applications in Dynamic Pricing of Retail Markets', in Proceedings of the 2003 IEEE International Conference on E-Commerce Technology (CEC '03), Pittsburgh, USA, pp. 339-346.

[11] C. V., Goldman, S. Kraus, & O., Shehory, 2004, 'On Experimental Equilibria Strategies for Selecting Seller and Satisfying Buyers', Decision Support Systems, vol. 38(3), Dec. 2004, pp. 329-346.

[12] M. Gupta, K. Ravikumar, & M. Kumar, 2002, 'Adaptive Strategies for Price Markdown in a MultiUnit Descending Price Auction: A Comparative Study', in Proceedings of the 2002 IEEE Conference on Systems, Man, and Cybernetics, Hammamet, Tunisia, pp. 373-378.