# 強化学習を用いたエージェントによる自動交渉システム

新井　成一† 　三浦　孝夫†

† 法政大学 工学研究科　〒 184–8584 東京都小金井市梶野町 3-7-2
E-mail: †07d3005@stu.hosei.ac.jp, ††miurat@k.hosei.ac.jp

**あらまし**　近年，インターネットの普及によりネットワーク上での新たな決済システムとして，電子商取引（オークションや為替取引）が利用されている．これにより，売り手と買い手との双方にとって，取引コストを高信頼のままに軽減することが可能となり資本の流動性を保ち円滑な経済活動を促すことができる．一層の改善，軽減を目的として，エージェント学習機構を利用した研究が注目を集めている．本研究では，商取引の場を環境とみなし，エージェントが環境に対する学習過程を通じて，利用者意思を反映し完結した取引を可能とする方式を提案する．

**キーワード**　電子商取引, オークション, 組合せオークション, 強化学習,Q 学習

# Automating Negotiation for Auction Based on Reinforcement Learning

Seiichi ARAI† and Takao MIURA†

† Dept.of Elect.& Elect. Engr., HOSEI University　3-7-2, KajinoCho, Koganei, Tokyo, 184–8584 Japan
E-mail: †07d3005@stu.hosei.ac.jp, ††miurat@k.hosei.ac.jp

**Abstract**　In this work, we discuss how to automate negotiation for auctions based on reinforcement learning. According to wide-spread Internet Technologies, we have many chances to see Electronic Commerce (EC) activities such as Auction and Exchange control as a new kind of payment. Both sellers and buyers allow to reduce transaction cost while keeping high reliability of market by which we can keep sound floating capitals and smooth economic activities. One of the issues of EC is how we negotiate market situation for the purpose of economic activities, in fact, we should keep watching the market all the time and decide what kinds of actions should be taken. In this investigation we propose automating negotiation process by means of agent technology with machine learning, and examine how well auction process works by some experimental results.

**Key words**　e-Commerce, Auction, Combinational Auction, Reinforcement Learning, Q-Learning

## 1. 前　書　き

近年，インターネットの普及により人々がネットワーク上で商取引を行う電子商取引 (EC:Electronic Commerce) を利用する機会が増えている．現在，この電子商取引の形態はオンラインショッピング (Amazon，楽天市場)，インターネットオークション (Yahoo!オークション，楽天オークション，e-Bay)，オンライントレード (ネット証券，FX トレード)，など多種多様である．これらオンライン上での商取引は，本来の商取引の形態に比べ，時間や場所の制約を軽減した商取引が行えることから社会的に高い関心を集めている．特にインターネットオークションは企業から個人まで様々な形態で取引が行われており，その利用者は世界中で増加している．

一方，この様な電子商取引において，人工知能分野におけるエージェントを用いた研究が注目を集めている. [7] エージェントとは，利用者の代理としてソフトウェア上で自律的に振舞う活動主体である．例えばエージェントは利用者の希望の商品を探し出したり，実際に交渉をする等のことを代理に行う．この様な研究では，エージェントの支援による利用者の取引コスト (Transaction Cost) の軽減が期待されている. [1] また更なる取引コストの削減を目指すべく，エージェント自身が決断する自動交渉システムの研究も注目されている. [6]

そこで本研究では，米国連邦通信委員会 (FCC) による周波数割り当てオークションの成功以来，ゲーム理論の発展形として注目を浴びている組合せオークションに対する自動交渉システムの実現を目的とする．

## 2. 複数ラウンド組合せオークション

組み合わせオークション (Combinatorial Auctions) とは．所与の財集合に対して入札者が同時一斉秘密に入札し．入札額の組み合わせが最大になるようなオークションをいう．入札は単一財ごとではなく．それらの組合わせを自由に対象としてよい．

表 1　勝者決定問題

| 入札者 | 出品財 | | | 入札額 |
|---|---|---|---|---|
| | パソコン | メモリ | HDD | |
| A | ○ | ○ | | 100 |
| B | | ○ | | 30 |
| C | | | ○ | 80 |

例えば, 組合せオークションで出品されている財の集合をパソコン，メモリ，HDD　とする．この時，入札者 A は パソコン，メモリ｝に金額 100 で入札をする．これは入札者 A はパソコン，メモリ｝の組合せに対しての評価額であり，またこの様な財を補完財という．評価額とは入札者の財に対する支払っても良いとする最大の金額のことである．入札者 B はメモリに対し金額 30 で入札をする．入札者 C は HDD に対し金額 80 で入札をする．これを表 1 に示す.

組合せ問題に対する解とは，入札額総和が最大となる衝突のない財割り当てをいう．解は入札結果として公開される．表 1 の例では，入札者 A,C の入札金額 180 が最大の組合せとなる．解を求める方式 (プロトコル) は, 英国式やビィックレー式入札方法が知られているが，ここでは最も単純な秘密英国式 (同時に入札し最高入札額が落札) を仮定する.

また組合せオークションにおける勝者決定問題は，入札者数を n とすると，勝者の可能な組合せの数が 2n と指数的に増加する．これは NP 完全と呼ばれる問題のクラスに属する．NP 完全な問題は多項式時間で解く方法は発見されておらず，最悪の場合には計算時間が n に関して指数的になるという性質がある.

組合せオークションにおける入札者にとっての利点は，単一では財の評価額を図れないものについての組合せ入札が可能という点にある．一方，出品者にとっての利点は一度に複数の財を出品できることによる取引コストの減少である.

本研究では, この組合せオークションを何度も適用し入札修正がなくなるまで繰返すという同時繰り返しオークションを扱う. [4] このオークションでは，各入札ラウンドは同時秘密に行われるが，入札者を除く割当て結果 (組み合わせと金額) がラウンドごとに公開され，入札者はこれを参考にして入札金額のみを修正 (増加) する．変更が無くなった時点でオークションは終了する．複数ラウンド組合せオークションのモデルを図 1 に示す.

## 3.　学習エージェント

以上のオークションモデルにおいて，参加者が得られる情報は相手からの提案のみであり，行動の基準を確立することは困難である．そこで，本研究では学習機能を利用したエージェントをオークションに参加者させ，エージェントがよりよい報酬の獲得を目指す方式を提案する.

本章では，組合せオークションにおいてエージェントが戦略的な行動を確立するための，学習機構について述べる．前章で提示したオークションモデルにおいて，各参加者が得られる情報は最大入札者，最大入札額，最大入札組合せのみである.

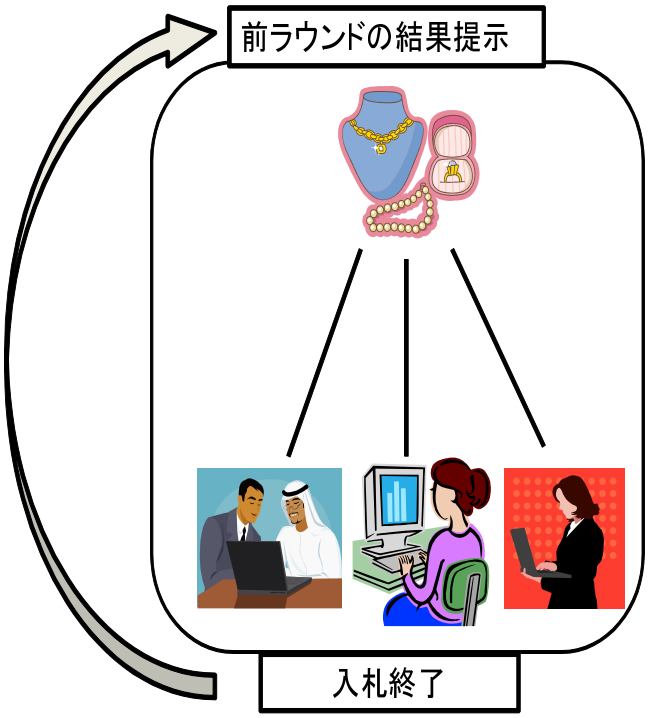


図 1　複数ラウンド組合せオークション

よってここではオークションを環境とみなし，環境から最適な行動を決定する学習手法として知られている Q 学習を適応する. [5] 提案するアルゴリズムの概要を図 2 に示す.

エージェントは受信部，学習機構，評価部，行動決定機構から構成される．受信部は環境から情報を受信する機構である．ここでの環境とは，オークション環境を指す．学習機構は過去の行動に対する結果のデータを保持し，最適な行動を学習する機構である．評価部は，利用者の嗜好を反映させた財への評価値を保持する機構である．行動決定機構は受信部，評価部，学習機構，から得られた情報を元に最終的な決定を行う機構である.

①エージェントは環境から情報を得る．ラウンド毎おけるにオークションの結果（最高入札者，最高入札組合せ，最高入札額）を受信する.
②受信部は行動決定機構に得た情報を送信する.
③行動決定機構は得られた情報をもとに，学習機構から過去の経験を要求する.
④そして，学習機構は過去の経験を送信する.
⑤また，評価部は評価値の情報を送信する.
⑥得られた情報を統合し，行動決定機構は最終的な行動を選択し，環境に対して送信する.
⑦ここで，前回の行動に対しての結果が環境から受信部へ送信される.
⑧エージェントはその結果を学習機構に送信し，(状態ー行動) ルールの価値を更新する.

以下同様に，学習を繰り返し行う．本研究で提案するアルゴリズムにおいて，図??の学習機構，行動決定機構はエージェントの根幹をなす部分である.

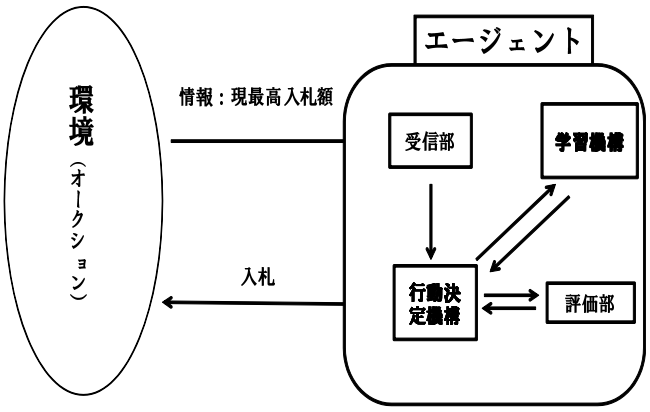


図 2　アルゴリズム概要

### 3.1 学習機構

組合せオークションにおいて，エージェントが得られる情報は最大入札者，最大入札額，最大入札組合せのみである．

この様な不完備な情報下では，エージェントは環境に対して試行錯誤を繰り返すことによって最適な行動を学習していく強化学習の適応が妥当と考えられる．そこで，本研究では強化学習の中でも有名な手法の一つであり，マルコフ決定過程の環境では学習率が適切に調整されれば，無限時間での学習の収束が保証されている Q 学習を適応する. [6] 以下に，その更新式を示す.

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha(r + \gamma \max_{a' \in A(s')} Q(s',a')) \tag{1}$$

この式において，$Q(s,a)$ は状態 s において行動 a をとるときの価値であり，$Q(s',a')$ は遷移先の状態 $s'$ において行動 $a'$ をとるときの価値を表している．また，$r$ は状態 s から $s'$ への遷移先において得られる報酬，$A(s')$ は状態 $s'$ で実行可能な行動全体を示し，$\alpha(0 < \alpha \leqq 1)$ は学習率，$\gamma(0 < \gamma \leqq 1)$ は割引率を表している．

エージェントは，ある状態から行動を選択する，という過程を繰り返し行うことで得られた結果をもとに Q 値を更新する．

### 3.2 行動決定機構

行動決定機構は，学習機構によって得られた Q 値，受信部から得られた情報，評価部から得られた情報，をもとに行動を決定する機構である．本研究では，エージェントの行動決定は Q 値に基ずく，$\varepsilon$ グリーディ手法を採用する．$\varepsilon$ グリーディ手法とは，確率 $1-\varepsilon$ でグリーディ手法を行い，確率 $\varepsilon(0 \leqq \varepsilon \leqq 1)$ でランダムな行動を選択する手法である．つまり，確率 $1-\varepsilon$ で Q 値が最大の行動をとり，確率 $\varepsilon$ でランダムな行動を選択する．この手法のポイントは，確率 $\varepsilon$ でグリーディ手法では更新されないルールの価値を更新できることである．

## 4. 組合せオークションにおける強化学習

西田らの英国式オークションにおける強化学習方式 [6] では，オークションの各状態・行動組み毎にメモリ空間を確保する．西田らは，”財への入札額”，”財への評価額と現在の入札額の差”，”入札行動”の 3 つを引数とした配列を用いる．学習のために，財に対しランダムに評価値が割り当てる．エージェントはその評価値を用いて，メモリ空間上の Q 値を基準とし入札行動をおこなう．与えられた評価値より低い財の落札額がエージェントの報酬となる (効用と言う)．エージェントはこの報酬を多くする行動を選択する．報酬を得たなら，エージェントは報酬を得た状態-行動に対して Q 値を更新する．これを終了すると，新たな組合せおよび評価値の下でこれを繰り返し学習する．

一方，本研究でもエージェントは各状態・行動 組みに対しメモリを確保する等，同様の学習を行う．すなわち，(各オークション毎に) 希望する財の組合せと評価値をランダムに割り当て，報酬を上昇させる入札行動を選択する．報酬を得た状態・行動の組に対して Q 値を更新し，これを繰り返し学習する．

組合せオークションでは，財の希望する組合せを割当てるとき，いくつかの引数を追加する必要がある．例えば，表 1 のパソコン，メモリ，HDD の財に対する組合せオークションでは，メモリ空間の確保を行うとき，配列の引数は，”財への希望組合せ”，”財への評価額”, ”入札行動” の 3 つとする．このとき ”財への希望組合せ” の引数に必要とされる空間数は，パソコン, メモリ,HDD, パソコン, メモリ, パソコン,HDD, メモリ,HDD, パソコン, メモリ,HDD の計 $2^3 - 1 = 7$ パターン存在する．つまり財の数を $m$ としたとき財の組合せ数は，$m^2 - 1$ 個存在し，扱う財の数を増やすに伴い入札の組合せのパターンが指数関数的に増大する．一方，組合せオークションでは各組合せパターンは独立であり予め制限を加えるわけにはいかないため，組合せオークションにおいて状態数の組合せ爆発は回避できない．

本研究にでは予め必要な状態数を静的に確保せず，ランダムアクセス可能なメモリ管理方を用いて，実際の要求の生じた状態・行動のみ Q 値を格納するという方法をとる．さらに，実験中に一定の間隔ごとに閾値以下の Q 値を削除を行う．この結果，ほとんど存在し得ないような状態・行動の除外や 微小は Q 値を有する状態・行動組みを削除することによって，メモリ空間の実質的節約および計算の高速化を計る．この手法は通常の Q 学習においても有用であるが，組合せオークションにおける状態数爆発の問題におけるこの削除効果は，現実にはかなり大きい．

## 5. 実　　験

本章では，提案したアルゴリズムの有用性を示すための実験方法を述べる．実験は第 2 章で述べた，複数ラウンド組合せオークションを適応する．

### 5.1 実験準備

オークションは入札額に変動が見られない，または制限ラウンドを過ぎた場合終了とする．また，1 回のオークションの終了までを 1 エピソードとし実験は計 100,000 エピソード繰り返し行う．さらに状態数の削減は 5000 エピソード毎に行い，削除する Q 値の閾値は 0 以下とする．また，オークション環境設定の一部として単純型の入札者をエージェントのほかに 49 名用意する．これらの入札は単純であり，評価値はランダムで与えられ入札はその評価値を制限としランダムに吊り上げを行う設定とする．

表 2　オークション設定

| 財の数 | 制限ラウンド | 参加者数 | 1 つの財に対する入札制限金額 |
|---|---|---|---|
| 3 | 10 | 50 | 100 |

表 3　エージェント設定

| 状態 | | | | 行動 |
|---|---|---|---|---|
| 入札組合せ | 評価値 | 現最高入札組合せ | 現最高入札額 | 入札額 |
| 7 | 0-300 | 5 | 0-300 | 0-300 |

本実験におけるオークションの初期設定は表 2 に示す．3 つの財はそれぞれ (A,B,C) とする．制限金額とは財 1 つに対する制限のことであり，2 つの財に対しての組合せ入札であれば制限金額は 200 となる．

本実験におけるエージェントの学習機構の設定を表 3 に示す．入札の組合せはオークションに出品される財によって変化する．本実験においては $2^3-1$ で 7 となる．評価値は 1 つの財に対し 100 を制限として乱数でオークション開始時に与えられる．現最高入札組合せは，オークションが終了しない限り前ラウンドの結果として与えられる．本実験においては　(A)(B)(C),(AB)(C),(A)(BC),(AC)(B),(ABC)　の 5 つのパターンのことである．現最高入札額は，オークションが終了しない限り前ラウンドの結果として与えられる．入札はエージェントの 1 つの財に対し 100 を制限とした入札金額のことである．

またそれぞれ $\alpha=0.1$,$\gamma=0.1$,$\varepsilon=0.1$ とし，報酬:r は効用 (評価値-入札額) で与えられる．これはエージェントが希望の財に対し，いくら安く落札できたかをエージェントの価値基準としている．

### 5.2 評価方法

本節では評価方法について述べる．

本研究では学習機能を利用して，エージェントが限られた時間内で自身の行動基準を確立することを目指している．ここでの行動基準とは効用ことを指す．つまり本実験では，エージェントが効用を保つ行動選択をすることを評価の基準とする．また，同時に Q 学習における必然性として，Q 値の収束も評価の基準とする．

### 5.3 実験結果

本節では実験結果を示す．図 3 は 5000 エピソードごとの平均効用を示している．エピソードを重ねるごとに効用が上昇しているのが分かる．図 4 は，状態が入札組合せ:(A)(B)(C), 評価値:150, 最大入札組合せ:(A)(B)(C) の時，入札:100 の行動を選択した場合の Q 値の遷移である．図 5 は，図 4 におけるエピソードごとの平均と分散を示している．図 6 は，各組合せにおける最大の Q 値である．

また状態の削減数は $800,000$(最大状態数) $-$ $13,164$(使用状態数) $=$ $1786,836$ となり，使用率は 1%以下と分かる．

### 5.4 実験考察

図 3 では効用の上昇が認められる．これは，エージェントがより評価値以下で財を落札している回数が増えていると考えられる．

| エピソード | 平均効用 | エピソード | 平均効用 |
|---|---|---|---|
| 5000 | 7 | 55000 | 10 |
| 10000 | 7 | 60000 | 13 |
| 15000 | 7 | 65000 | 12 |
| 20000 | 8 | 70000 | 12 |
| 25000 | 7 | 75000 | 12 |
| 30000 | 7 | 80000 | 15 |
| 35000 | 8 | 85000 | 14 |
| 40000 | 9 | 90000 | 15 |
| 45000 | 10 | 95000 | 22 |
| 50000 | 9 | 100000 | 20 |

図 3　5000 エピソードごとの平均効用

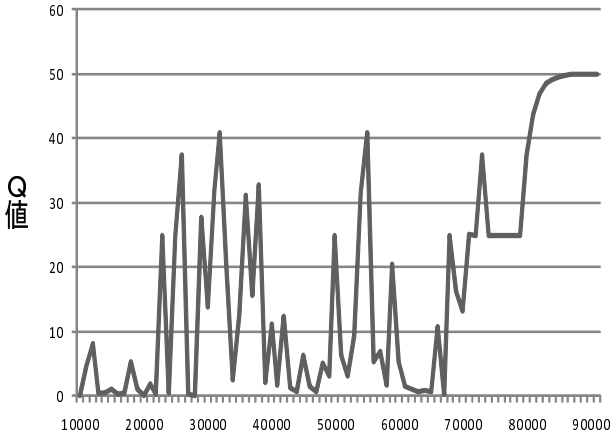


図 4　Q 値の収束

| エピソード | 1000 | 2000 | 3000 | 4000 | 50000 | 60000 | 70000 | 80000 |
|---|---|---|---|---|---|---|---|---|
| 平均 | 2.26 | 20.91 | 19.12 | 4.6 | 15.39 | 6.25 | 26.35 | 47.48 |
| 分散 | 9.45 | 359.34 | 182.73 | 22.41 | 222.51 | 71.71 | 17.34 | 16.12 |

図 5　Q 値の平均分散

| 希望組合せ | 評価値 | 現最高評価値 | 現最高組合せ | 行動 | Q |
|---|---|---|---|---|---|
| A | 99 | 219 | (A),(BC) | 4 | 63.75 |
| A | 94 | 278 | (A),(B),(C) | 11 | 46.5 |
| B | 96 | 276 | (A),(B),(C) | 12 | 42 |
| B | 97 | 278 | (A),(B),(C) | 17 | 40 |
| C | 83 | 269 | (AB),(C) | 10 | 62 |
| C | 96 | 278 | (AB),(C) | 10 | 48 |
| AB | 195 | 255 | (AB),(C) | 5 | 95 |
| AB | 191 | 250 | (AB),(C) | 12 | 94.5 |
| AC | 114 | 240 | (AC,(B) | 72 | 21 |
| AC | 120 | 274 | (AC),(B) | 110 | 5 |
| BC | 183 | 288 | (A),(BC) | 10 | 92 |
| BC | 185 | 266 | (A),(BC | 12 | 92 |
| ABC | 283 | 204 | ABC | 212 | 31 |
| ABC | 280 | 232 | ABC | 235 | 23 |

図 6　各組合せにおける最大の Q 値

図 4 では 80,000 エピソード付近で Q 値の収束が確認できる．さらに Q 値は．図 4 の状態における最大理想効用値:50 へと収束していくのが確認できる．また図 5 より，Q 値の平均値は上

昇しており，分散は低下していくことからも Q 値の収束が確認できる．

図 4 では，エージェントの希望入札組合せと現最高入札組合せとの区分けが適している場合に最大の Q 値を得ていることが分かる．これは，組合せオークションの特性から当然の結果だと考えられる．

また最終的に使用した状態数は 1%以下であり，ほとんど不要な状態なのだと分かる．これは評価値以上の入札をしても得られる報酬は負であることや，現最高入札額が当然ある一定以上の値になることが理由だと考えられる．

## 6. 結　　論

本研究では組合せオークションににおいて，強化学習を用いたエージェントの自動交渉システムを提案した．本手法により，オークションを環境とみなすことでエージェントが戦略的な行動を獲得したことを実験により示した．また，組合せオークションの研究において必然となる状態数の増加に対する対処法を示した．

## 文　　献


[1] Robert H.Guttman and Pattie Maes: Cooperative vs. Commpetitive Multi-Agent Negotiations in Retail Electronic Commerce, Cooperative Information Agents II Learning, Mobility and Electronic Commerce for Information Discovery on the Internet Lecture Notes in Computer Science, 1998, Volume 1435/1998, 135.

[2] Lucian Busoniu,Robert Babuska and Bart De Schutter: Multi-Agent Reinforcement Learning: A Survey. Control, Automation, Robotics and Vision, 2006. ICARCV '06. 9th International Conference on Issue Date: 5-8 Dec. 2006 ,On page(s): 1 - 6 ,Location: Singapore ,Print ISBN: 1-4244-0341-3 ,INSPEC Accession Number: 9484363.

[3] Liping Fang and Yucheng Wang: OICAS:An Online Interative Combinatorial Auction System. 2005 IEEE International Conference on Issue Date: 10-12 Oct. 2005,On page(s): 233 - 238 Vol. 1.

[4] Leyton-Brown, K.,Shoham,Y. ,Tennenholtz,M.: An algorithm for multi-unit combinatorial auctions, National Conference on Artificial Intelligence,2000

[5] 高玉　圭樹: マルチエージェント学習, コロナ社,2004.

[6] 大竹麗尾, 西田豊明: 強化学習を用いた交渉戦略アルゴリズム, 電子情報学会,OFS2001-11,AI2001-16(2001-07).

[7] 伊藤孝之, 服部宏光, 新谷虎松:エージェント間の協調的入札機構に基づく複数オークション入札支援システム BiddingBot, 人工知能学会誌,17-3-a(2002)