JIT2014

Identificación del Trabajo	
Área:	Electrónica, informática y comunicaciones
Categoría:	Alumno
Regional:	Santa Fe

Agentes inteligentes en mercados bilaterales

Constanza QUAGLIA

Centro de Investigación y Desarrollo en Ingeniería en Sistemas de información CIDISI (Lavaisse 610, Santa Fe), Facultad Regional Santa Fe, UTN

E-mail de contacto: cotii.q@gmail.com

Este trabajo ha sido realizado bajo la dirección de la Dra. Mercedes Canavesio, en el marco del proyecto "Arquitectura cognitiva multi-agente para control distribuido, scheduling emergente, supervisión y optimización autonómica de sistemas de producción".

Resumen

Las empresas enfrentan profundos cambios en su entorno de negocios, que las conducen a formar alianzas estratégicas con sus pares para satisfacer demandas de mercado más exigentes. Se propone un modelo de compañía fractal, para la integración virtual y temporal entre empresas, cuya estructura organizacional se sustenta *relaciones cliente-servidor*, establecidas entre gestores de proyectos. En este modelo de *mercado bilateral* existen clientes que anuncian requerimientos por recursos a ciertos proveedores, y estos, seleccionan a un conjunto de clientes a quienes desean proveerles sus recursos. Lograr un par cliente-servidor efectivo, es un problema no determinístico y complejo. Por ello, se propone incorporar en los gestores de proyecto y de recursos capacidad para aprender, mediante algoritmos de *aprendizaje por refuerzo*, enfocándolo en el problema del *bandido de 2-brazos.*

Palabras Claves: Aprendizaje por Refuerzo; Mercados Bilaterales; Problema bandidos 2-brazos

1. Introducción y Objetivos

Las empresas se enfrentan a profundos cambios en su entorno de negocios, que las conducen a formar alianzas estratégicas con sus pares para así satisfacer demandas de un mercado, cada vez más exigente. Para que las empresas que participan de estas redes o alianzas alcancen los beneficios y ventajas competitivas esperadas, se requiere desarrollar un modelo de empresa integrada que identifique y defina concretamente la estructura, procesos, información y relaciones entre las empresas que las componen. Atendiendo a esta problemática, Canavesio y Martinez (2007) proponen un modelo de compañía fractal basada en proyectos para la integración virtual y temporal entre empresas. En este modelo, la unidad fractal de gestión es el proyecto, el cual es una entidad auto-gestionada, interdependiente y temporal, que combina distintos tipos de habilidades, conocimientos y recursos para lograr una meta concreta (ej. completar una orden, diseñar un nuevo producto, satisfacer un requerimiento por recurso, etc.).

La unidad fractal de gestión propuesta está compuesta de un gestor de proyecto que gestiona la misma y un objeto que es gestionado por éste. Dado que en el modelo, tanto los fines o metas como los recursos o medios son gestionados a través de proyectos, el objeto gestionado

es la meta del proyecto que está asociada con el logro de una dada resultante (fin) o la prestación de un dado recurso (medio). Por ello, el gestor de un proyecto asumirá el rol de gestor de fines o gestor de medios, respectivamente. El gestor de un proyecto es un agente inteligente capaz de actuar de manera autónoma para alcanzar sus objetivos, de aprender capitalizando experiencia, modificando su política de actuación y su representación interna del entorno para adaptar su comportamiento a los cambios del mismo.

El modelo de compañía fractal propuesto es un mercado de coincidencias (Sotomayor, 2004) donde existen clientes que anuncian requerimientos por recursos a diversos proveedores (servidores), y a la vez, existen proveedores de recursos que seleccionan a un conjunto de clientes a quienes desean proveerles sus recursos. Lograr que un dado cliente concuerde con un dado servidor para que satisfaga un requerimiento determinado, requiere que los gestores de proyecto conozcan preferencias, habilidades y estrategias de los otros gestores con los que negocian y/o compiten. Cuando un agente cliente y un agente servidor de recursos llegan a un acuerdo, entre ellos se establece una relación cliente-servidor a través de la cual interactúan. Este concepto de relación cliente-servidor, es fundamentalmente importante para el modelo de empresa integrada. Por ello, es necesario dotar a los gestores de proyecto con capacidad de aprendizaje que les permitan conocer su entorno, sus competidores y servidores para realizar negociaciones entre agentes más beneficiosas económicamente.

En este marco, el objetivo general de este trabajo es analizar e implementar algoritmos de aprendizaje por refuerzo abordando en particular, el problema de los 2-bandidos, dado que es el modelo más apropiado para aplicar en mercados de coincidencias como es el caso de la compañía fractal basada en proyectos. De esta manera los gestores de proyecto aprenden a conocer su entorno y establecer relaciones cliente-servidor sólo con quienes les convienen económicamente.

2. Metodología

2.1. El modelo de la compañía fractal basada en proyectos

La idea de la compañía fractal (Warnecke, 1993) es un modelo de empresa conceptual, que a través de unidades autónomas, descentralizadas e interdependientes, denominadas fractales, otorga a las empresas mayor flexibilidad y agilidad para adaptarse a los cambios en su entorno de negocios. Un fractal es definido como una estructura que describe un patrón idéntico, que se replica a sí mismo a distintos niveles de abstracción, de manera recursiva. En el modelo de empresa fractal propuesto por Canavesio y Martinez (2007), la unidad fractal de gestión se concibe como un proyecto. Esta unidad fractal de gestión o proyecto combina distintas habilidades, destrezas y competencias necesarias para lograr un objetivo específico (satisfacer un requerimiento por recursos, diseñar de un nuevo equipo de RX, desarrollar una prótesis, etc.).

Dentro de la red de empresas, cada proyecto posee las siguientes propiedades que están presentes en cada instancia, más allá de su nivel de abstracción: i) Orientado a la meta: cada proyecto produce una resultante específica para un cliente claramente definido. ii) Autónomo: cada gestor de proyecto posee suficiente libertad para ejecutar actividades y gestionar los recursos involucrados para el logro de una dada meta. iii) Temporal: cada proyecto debe lograr su meta en una cantidad limitada de tiempo. iv) Relación recursiva: cada proyecto puede ser definido como parte de un super-proyecto o puede contener distintos niveles de sub-proyectos (sub-proyectos, sub-sub-proyectos, …, etc.) v)Ciclo de vida que cubre cuatro etapas: definición,

ejecución, control y cierre.vi) La información y conocimiento relevante del proyecto es almacenada y compartida por todos los gestores intervinientes y les permiten mejorar sus futuras decisiones.

En el modelo de la compañía fractal basada en proyectos, la unidad fractal propuesta se compone de un gestor de proyecto que gestiona la misma y de un objeto que es gestionado por éste (Figura 1). El gestor de proyectos es un actor o agente inteligente, que posee suficiente libertad para tomar decisiones, ejecutar acciones, aprender y ajustar permanentemente su comportamiento. Dado que en el modelo, tanto los fines o metas como los medios o recursos son gestionados a través de proyectos, el gestor de un proyecto asumirá el rol de gestor de fines o gestor de medios, respectivamente. Ambos roles se establecen con funciones y responsabilidades claramente definidas.

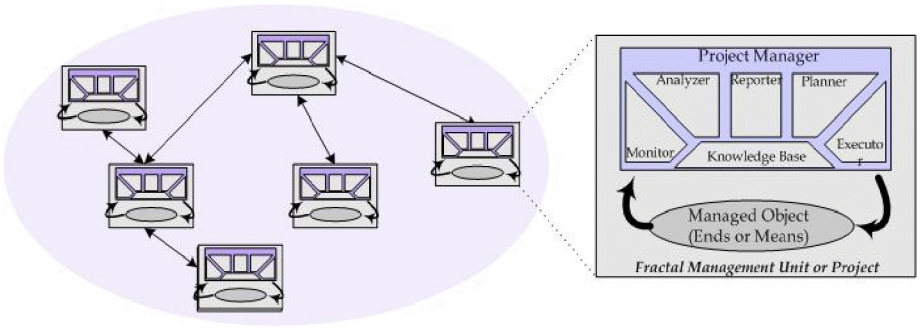

Figura 1. Estructura del proyecto como unidad fractal de gestión

El modelo de compañía fractal propuesto es un mercado de coincidencias (Sotomayor, 2004) donde existen clientes que anuncian requerimientos por la provisión de recursos a diversos proveedores, y a la vez, existen proveedores de recursos que seleccionan a un conjunto de clientes a quienes desean proveerle sus recursos. Las empresas se vinculan entre sí a través de relaciones cliente-servidor, entre gestores de fines (clientes) y gestores de medios (servidores), que pueden pertenecer o no a la misma compañía. Estas relaciones se establecen a través de algún mecanismo de negociación entre agentes interesados. Así, la compañía fractal se ve como un conjunto de relaciones temporales cliente-servidor, a través de las cuales los gestores de proyecto interactúan para diversificar su portafolio de productos, acceder a una mayor variedad de recursos, reducir costos, tiempo e incertidumbre.

2.2. El aprendizaje por refuerzo en los agentes que actúan en mercados bilaterales

El modelo de la compañía fractal propuesto es un mercado bilateral (two-sidedmarkets) (Sarne y Kraus, 2008; Rochet y Tirole, 2008; Kumar y otros, 2010; Chen y Song, 2013), donde existen clientes que anuncian requerimientos por la provisión de recursos a diversos proveedores, y a la vez, existen proveedores de recursos que seleccionan a un conjunto de clientes a quienes desean proveerle sus recursos. Así se logran pares cliente con servidores de recursos que definen una relación cliente-servidor, por lo que el modelo se debe considerar como un *mercado de coincidencias* (Sotomayor, 2004). Para ello, a través de la incorporación de algoritmos de aprendizaje por refuerzo en ambos roles de los gestores de proyecto, y enfocándolo en el

problema del bandido (two-sided bandit problem) que les permitirá hallar coincidencias entre ellos, se propone para lograr el objetivo de este proyecto.

El aprendizaje por refuerzo (Sutton y Barto, 1998) es un enfoque computacional para entender y automatizar el aprendizaje orientado al logro de metas y toma de decisiones en una secuencia. Mientras un agente interactúa con su entorno, aprende por prueba y error cual acción ejecutar. En cada episodio, el agente selecciona una acción posible en el actual estado y la ejecuta, causando que el entorno se mueva al siguiente estado. El agente recibe una recompensa que refleja el valor de la acción tomada. El objetivo del agente es maximizar la suma de las recompensas acumuladas desde un estado inicial hasta que alcanza el estado final. Inicialmente, el agente desconoce el curso de acción a tomar en función del contexto. A través de la interacción, el agente descubre qué acciones tienen mayor recompensa tras un análisis retrospectivo de los resultados (aciertos y errores) que ha obtenido. La implementación de agentes que aprenden por refuerzo se lleva a cabo utilizando una estructura compuesta por los siguientes elementos (Sutton y Barto, 1998):

- Política, define el objeto de optimización y mejora el conocimiento disponible por el agente.
- Función recompensa, define el objetivo que se espera satisfacer al final de cada episodio.
- Función valor o utilidad proporciona una medida de la efectividad de una política dada.
- El modelo del entorno imita el comportamiento del mismo.

En el caso particular del *problema de bandido de n-brazos (*Sutton y Barto, 1998), un agente debe elegir cuál de los n brazos jalar en cada período de tiempo para maximizar la recompensa recibida, mientras simultáneamente trata de estimar la distribución de recompensas de cada uno de los brazos. El agente debe decidir entre jalar el brazo con el valor más alto esperado y jalar el brazo que le permita aprender más sobre su distribución de recompensas. Se propone especializar esta situación al problema del bandido de 2-brazos, donde un brazo obtiene una recompensa basado en quien lo jaló y que él puede rechazar a quien lo hizo.

El problema de aprendizaje enfocado como el problema del bandido de 2-brazos es una formulación natural para mercados en los cuales existen dos tipos diferentes de agentes que deben lograr coincidencias entre ellos, repetidas veces. Ejemplos de estos mercados son el de citas, en el cual hombres y mujeres van en varias ocasiones a citas mientras aprenden sobre sus candidatos; el mercado laboral, en donde empleadores y potenciales empleados aprenden el uno del otro durante las entrevistas (Rochet y Tirole, 2006; Das y otro, 2005; Das, 2006), como así también el mercado de la compañía fractal basada en proyectos, cuando los gestores de proyecto deben decidir con quién asociarse para establecer relaciones cliente-servidor. Si bien un simple problema del bandido de 2-brazos no podrá capturar todos los aspectos de estos mercados si puede proveer un útil punto de partida para estudiarlos.

3. Resultados y discusión

3.1. El modelo de aprendizaje

Como se ha descripto en apartados previos, el modelo de la compañía fractal define dos roles para los agentes: gestores de fines y gestores de medios. Cada agente pertenece a un solo tipo y puede relacionarse únicamente con alguno del otro tipo. Existen *F* gestores de fines y *M* gestores de medios, que interactúan durante *T* períodos de tiempo o episodios intentando aprender y conocer sus preferencias en cuanto a socios para establecer relaciones cliente-servidor.

3.1.1. Mecanismo de emparejamiento

El mecanismo de macheo está basado en el algoritmo de Gale-Shapley Das y Kamenica (2005), complementado con la técnica de Q-learning (Sutton y Barton, 1998). En el algoritmo de Gale-Shapley, cada agente conoce a priori sus preferencias con respecto a los demás agentes. Un procedimiento centralizado se encarga de generar las asociaciones entre agentes en base a esas preferencias.

No obstante, en este modelo los agentes en principio no cuentan con información suficiente sobre los demás por lo que todos tienen el mismo nivel de preferencia respecto al resto. A medida que interactúan, irán aprendiendo de ellos y definiendo de este modo sus preferencias. Esto implica que en lugar de tener una lista de preferencias, cada agente tendrá una lista de valores Q, donde Q_i está asociado al gestor i del otro tipo. Al finalizar cada episodio, los gestores actualizan estos valores en función del resultado de la interacción realizada.

3.1.2. Decisión del gestor de fines

Como se mencionó anteriormente, cada gestor de fines tiene una lista de valores Q, donde Q_j está asociado al gestor de medios j. Inicialmente este valor es 0 para todos los gestores.

Cada episodio, el gestor de fines elige a un gestor de medios, según una política ε-greedy de selección de proveedor. Es decir que el agente puede tomar dos alternativas posibles: *explorar*, con una probabilidad asociada ε o *explotar*, con una probabilidad de 1-ε. Si decide *explorar*, entonces elegirá un gestor de medios al azar. En el caso de que decida *explotar*, elegirá aquel que tenga el máximo valor de Q. Finalmente, el agente le envía una solicitud de asociación al gestor de medios seleccionado.

3.1.3. Decisión del gestor de medios.

Al igual que los gestores de fines, cada gestor de medios tiene una lista de valores, donde Q_j está asociado al gestor de fines *j*. Inicialmente este valor se establece en 100 para todos los gestores.

En cada iteración, la decisión del gestor de medios está limitada a la elección entre aquellos gestores de fines que lo hayan elegido a él.

Cuando un gestor de fines *i* elige a un gestor de medios, pueden darse varias situaciones:

1. El gestor de medios está sólo (no está asociado con nadie).
2. El gestor de medios está asociado a otro gestor de fines *j* tal que $Q_i > Q_j$, es decir que el valor de Q para el que le propone supera al de su socio actual.
3. El gestor de medios está asociado a otro gestor de fines *j* tal que $Q_i \leq Q_j$, es decir que el valor de Q del que propone no supera al de su socio actual.
4. El gestor ya estaba asociado con i.

A estas alternativas se le suma la política de ε-greedy, con la cual el agente puede explorar o explotar. Finalmente, la decisión se toma de la siguiente manera:

1. Si el agente está solo, acepta.
2. Si no, aplica ε-greedy:
    a. Si decide explorar, acepta.
    b. Si decide explotar, acepta si la nueva propuesta tiene mejor valor de Q que su socio actual (caso 2) o si ya estaba asociado con ese mismo gestor (caso 4), y rechaza si la nueva propuesta tiene peor Q que su actual pareja (caso 3).

3.1.4. Actualización de las variables.

Al final de cada episodio los gestores actualizan sus variables internas. El gestor de fines actualiza su valor de Q correspondiente al gestor de medios que eligió, según la ecuación (1), donde r_{t+1} es la recompensa obtenida de dicha asociación y B_{t+1} es el factor de credibilidad.

Este factor evalúa cuán bien el servidor se desempeñó. Para los restantes gestores de medios mantiene el valor de *Q* anterior.

$$Q_{t+1} = Q_t + \alpha[r_{t+1} - Q_t + B_{t+1}] \quad (1)$$

El factor de credibilidad se establece inicialmente en 100 para todos los gestores, dado que los considera a todos igualmente confiables y capaces de satisfacer sus requerimientos. Luego, en cada episodio se calcula el mismo de acuerdo con la ecuación (2).

$$\mathrm{B_{t+1}} = \begin{cases} \mathrm{B_t} & \text{proveedor capaz} \\ \mathrm{B_t} - 25 & \text{proveedor incapaz} \end{cases} \quad (2)$$

Se considera proveedor capaz a aquel que acepte la solicitud de asociación y además sea capaz de proveer efectivamente el recurso solicitado. De esta manera se busca que los gestores aprendan a elegir entre aquellos que son capaces de proveer sus necesidades y a descartar aquellos que no, asignándoles un valor de *Q* más bajo.

Por otro lado, cada gestor de medios actualiza el valor de *Q* del gestor de fines que lo eligió según las ecuaciones (3), (4) y (5). En la ecuación (4) *B* representa la recompensa obtenida y *P* es una penalidad que se define en la ecuación (5). Esta penalidad permite que el gestor de medios aprenda a aceptar aquellos contratos para el cual es capaz de proveer los recursos.

$$Q_{t+1} = Q_t + \alpha[r_{t+1} - Q_t] \quad (2)$$

$$r_t = B_t - P_t \quad (3)$$

$$P_{t+1} = \begin{cases} P_t & si\ el\ agente\ es\ capaz\ de\ proveer\ el\ recurso \\ P_t + 25 & si\ el\ agente\ es\ incapaz\ de\ proveer\ el\ recurso \end{cases} \quad (4)$$

3.2. Resultados preliminares

3.2.1. Los gestores de fines aprenden a seleccionar el mejor proveedor.

Se considera un escenario donde hay 2 gestores de fines y 5 gestores de medios. Los valores de recompensas son obtenidos de una distribución normal con una media de 60 y desviación estándar 20. La capacidad de los gestores de medios se inicializa aleatoriamente en 0 (no capaz) o en 1 (capaz). De aquí resulta que los gestores de medios 1 y 4 son capaces de proveer sólo al gestor de fines 2, el gestor de medios 5 es capaz de proveer sólo al gestor de fines 1, el 2 es capaz de proveer a ambos y el 3 es incapaz de proveer a ambos. Este escenario se simuló durante 1000 episodios.

La figura 2 muestra para el gestor de fines 1 el porcentaje de veces que eligió a cada gestor de medios como proveedor. Se observa que en un principio, el gestor comienza eligiendo al gestor de medios 1 aproximadamente un 94% de las veces. Después de unas 15 iteraciones este porcentaje decae, dado que el gestor de fines aprende que este es un gestor de recursos incapaz. Luego comienza a elegir al proveedor 2. Hasta la iteración 120 aproximadamente esta preferencia aumenta hasta un 80% y luego decae. Finalmente muestra una preferencia por el gestor de medios 5, preferencia que aumenta hasta el final de la corrida, con un valor de 80% en la última

iteración. Estos últimos dos gestores, 2 y 5, son capaces de proveer de los recursos que el gestor de fines necesita. Con el paso de las iteraciones, descubre que el proveedor de recurso identificado como 5, tiene un valor de recompensa mayor, debido a que responde y satisface más eficazmente sus requerimientos.

Del mismo modo, el gestor de fines 2 debe elegir entre los gestores de medios 1, 2 y 4. Finalmente se queda con el 1, ya que es el que mayor recompensa le da de los tres.

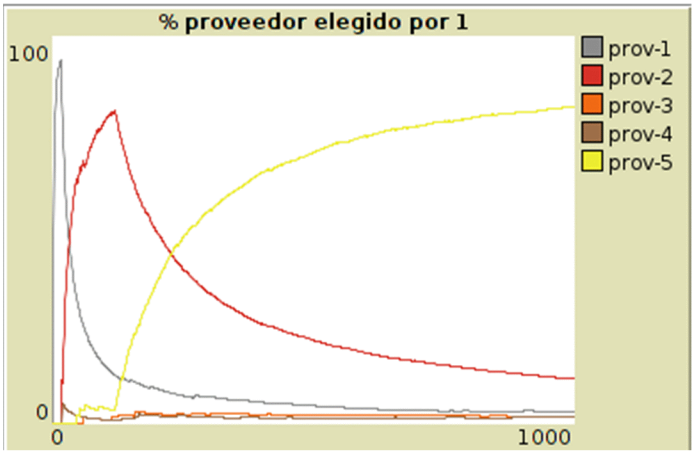

Figura 2. Curva de aprendizaje del gestor de fines 1.

3.2.2. Los gestores de medios aprenden a aceptar los mejores contratos

Se simula un escenario similar al del punto anterior, sólo que con 5 gestores de fines y 2 gestores de medios. El gestor de medios 1 es capaz de proveer a todos excepto al gestor de fines 4 y su mayor recompensa la obtiene del gestor de fines 5. El gestor de medios 2 se reconoce capaz de prestar recursos a todos los gestores de fines excepto al 3, y obtiene su mayor recompensa del gestor de medios identificado como 1. La figura 3 muestra el porcentaje de gestores de fines aceptados por el gestor de medios 1. Se puede ver su tendencia a aceptar al gestor de fines 5, con el cual maximiza su recompensa. Esta preferencia va siempre en aumento. Al final de la corrida, en la iteración 1000, eligió al gestor 5 en un 85% de las veces. De manera similar, el gestor de medios 2 acepta al gestor de fines 1 un mayor porcentaje de veces.

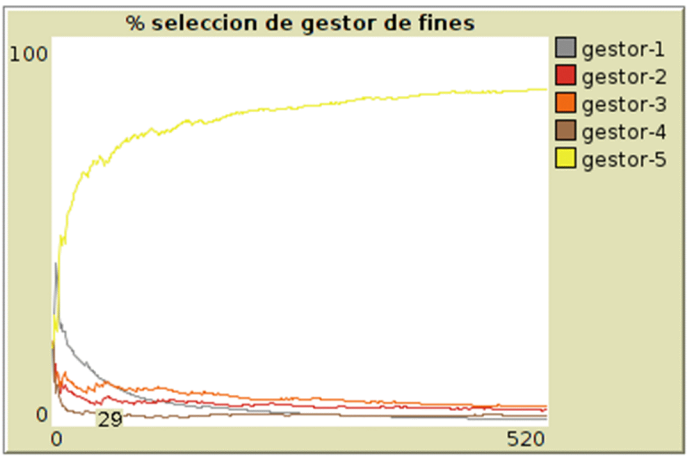

Figura 3. Curva de aprendizaje del gestor de medios 1.

4. Conclusiones

El trabajo presenta brevemente la descripción del modelo de la compañía fractal, que favorece la integración entre empresas temporal y virtual, en el logro de objetivos específicos de negocios. Las relaciones cliente-servidor entre gestores de proyectos, son el estructura fundamental el modelo de empresa, por ello, la selección de los socios es fundamentalmente importante para el éxito de la compañía fractal.

Las simulaciones realizadas sobre la incorporación de aprendizaje en los agentes de gestión para establecer relaciones cliente-servidor, mostró cómo los agentes clientes aprenden a requerir recursos sólo a proveedores confiables, que poseen la capacidad de proveer de recursos con la calidad requerida por la tarea y además de mínimo costo. Por otro lado, los agentes que se desempeñan como gestores de recursos aprenden a identificar a qué tareas son capaces de proveer un dado recurso para maximizar sus beneficios y credibilidad ante sus clientes.

Si bien el trabajo se basa en un modelo de la realidad sencillo, esta simplificación nos provee un útil punto de partida para estudiar la incidencia del aprendizaje en los agentes gestores de proyecto. La actual utilización de variables de recompensa y capacidad podrían extenderse para abarcar aspectos como beneficio económico de las interacciones, restricciones de tiempo de entregas, restricciones de calidad, etc.

Bibliografía

Canavesio, M, Martinez, E (2007) Enterprisemodeling of a Project-oriented fractal company for SMEs networking. Computer in Industry Nro 58 Pp 794-813.

Sotomayor, M.(2004) Implementation in the many-to-many matching market. Games and economic behaviour.Nro 46, Pp. 199-212.

Warnecke, H.J. (1993) The fractal campany. A revolution incorporate culture. Springer-Varlag. Berlin.

Sarne,D., Kraus, S, (2008) Managing parallel inquiries in agents` two-sided search. Artificial intelligent Vol 172 (4-5) Pp 541-569

Rochet,J, Tirole, J. (2008) Tying in two-sided markets and the honor all cards rule. International journal of industrial organization Vol 26 Nro 6 Pp 1333-1347.

Kumar,R., Lifshits,Y., Tomkins,A. (2010) Evolution of two-sided markets. ACM Proceeding of the third ACM international Conference on web search and dataming. ISBN 978-1-60558889-6 PP. 311-320

Chen,J., Song, K., (2013) Two-sided matching in the loan market. International journal of industrial organization Nro 33 Pp. 145-152.

Sutton, R. Barto,A (1998) Reinforcement learning. An introduction MIT Press.

Rochet,J, Tirole, J. (2006). Two-sided markets: a progress report. The RAND journal of economics Vol 37 (3) Pp 645-667.

Das,S., Kamenica,E., (2005) Proceeding 19th Intenational Joint Conference on Artificial Intelligence IJCAI'05. Pp. 947-952.

Das,S., (2006). Dealers, Insiders, and Bandits: learning and its effects on market outcomes. PhD Thesis.Massachusetts Institute of Technology.