# Agentes que aprenden a establecer relaciones cliente-servidor en mercados bilaterales

Constanza Quaglia[1]

[1] CIDISI Centro de Investigación y Desarrollo en Ingeniería en Sistemas de Información. Facultad Regional Santa Fe – Universidad Tecnológica Nacional (FRSF – UTN)

Constanza.iquaglia@gmail.com

**Resumen.** En respuesta a presiones competitivas, pequeñas y medianas empresas han comenzado a formar redes colaborativas entre ellas. En este trabajo se presenta un modelo de compañía fractal basada en proyectos, en el cual los gestores de proyecto establecen relaciones del tipo cliente-servidor para negociar la asignación de un determinado recurso. Se propone la incorporación de algoritmos de aprendizaje por refuerzo a los gestores, de manera que puedan aprender a seleccionar a su socio más conveniente a lo largo del tiempo, y así maximizar sus beneficios.

## 1 Introducción

La globalización de los mercados, los avances tecnológicos y la creciente personalización de productos y servicios están creando una fuerte competencia entre las empresas. Estas condiciones se acentúan en compañías dedicadas al diseño de productos y tecnologías innovadoras.

Para satisfacer las demandas de este mercado, cada vez más exigente, las empresas, sobre todo pequeñas y medianas, junto con organizaciones públicas y privadas, universidades y centros de investigación, están formando redes de asociación estratégicas.

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 [1] 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 y un objeto gestionado por éste. 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, asociada con el logro de una resultante (fin) o la prestación de un recur-

so (medio). Por ello, el gestor de un proyecto asumirá el rol de gestor de fines o gestor de medios, respectivamente. El modelo propuesto es un mercado de coincidencias [2], donde existen clientes que anuncian requerimientos por recursos y proveedores de tales recursos (servidores). Cuando un agente cliente y un agente servidor de recursos llegan a un acuerdo, entre ellos se establece una relación temporal cliente-servidor a través de la cual interactúan. Este concepto de relación cliente-servidor, es fundamentalmente importante para el modelo de empresa integrada.

En este marco, el objetivo general de este trabajo es analizar e implementar algoritmos de aprendizaje por refuerzo, de manera que los gestores de proyecto aprenden a conocer su entorno y establecer relaciones cliente-servidor sólo con quienes les convienen económicamente. Para ello, se aborda 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.

En [3] se presenta un algoritmo de aprendizaje en el cual los gestores aprenden a seleccionar al mejor socio de un conjunto de proveedores, al tiempo que los proveedores aprenden a aceptar sólo aquellas relaciones más convenientes. El presente trabajo extiende el mencionado, al incorporar la temporalidad intrínseca que cada tarea de un plan de proyecto posee. Así, la elección del socio se lleva cabo para cada tarea del plan que inicia en cada iteración, y el aprendizaje se centra no solo en la selección del socio sino más específicamente en conocer la capacidad del gestor de medios en relación a la tarea para la cual compromete recursos.

## 2 Contexto

### 2.1 Modelo de compañía fractal

La idea de la compañía fractal [4] 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 la unidad fractal de gestión se concibe como un proyecto. Dentro de la red de empresas, cada proyecto 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 de recursos, etc.).

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, respecti-

vamente. Ambos roles se establecen con funciones y responsabilidades claramente definidas.

El modelo de compañía fractal propuesto es un mercado de coincidencias 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 Aprendizaje por refuerzo

El modelo de la compañía fractal propuesto es un mercado bilateral (two-sided markets) [5, 6, 7, 8], 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. Para ello, se incorpora algoritmos de aprendizaje por refuerzo en ambos roles de los gestores de proyecto, y enfocándolo en el problema del bandido (two-sided bandit problem) les permitirá hallar coincidencias entre ellos.

El aprendizaje por refuerzo [9] 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:

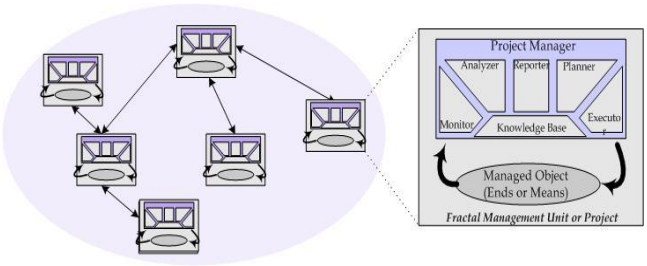


**Figura 1.** Estructura del proyecto como unidad fractal de gestión [1].

- *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, 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 el brazo que le permita aprender más sobre su distribución de recompensas. Así, 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 [10, 11, 12] 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 sí puede proveer un útil punto de partida para estudiarlos.

## 3 Modelo de aprendizaje

En la compañía fractal existen gestores de fines y gestores de medios, que interactúan durante una cierta cantidad de períodos de tiempo o episodios intentando aprender y conocer la confiabilidad y capacidad de los socios entre quienes establecer relaciones cliente-servidor exitosas.

Un plan de proyecto está integrado por cientos o miles de tareas que requieren de diferentes recursos y habilidades para su ejecución. El gestor de proyecto dispone de un centro de potenciales proveedores, algunos calificados como confiables y capaces de proveer recursos con el nivel de calidad requerido y otros que no reúnen tales condiciones. A priori el gestor de proyecto desconoce esta calificación de los potenciales proveedores por lo que, deberá aprender a distinguirlos. Para ello, utiliza un parámetro denominado *factor de credibilidad*, que considera la capacidad de un proveedor de cumplir en tiempo y forma sus contratos. Inicialmente, todos los gestores de recursos son considerados confiables para proveer de recursos para la tarea *ta*. Mientras el gestor de proyecto tenga poca experiencia sobre la credibilidad de los proveedores, incurrirá en elevados costos, por elegir servidores económicos e ineficaces o eficaces

pero onerosos. Cuando el agente ha adquirido conocimiento respecto a quienes son los servidores confiables y además capaces de proveerle los mejores recursos para una dada tarea, se reduce la búsqueda, la incertidumbre y los costos asociados debido a que las negociaciones se circunscriben sólo a este grupo de proveedores. Por otro lado, los proveedores de recursos deben aprender qué contratos de provisión aceptar para incrementar beneficios y credibilidad ante sus clientes. Durante su aprendizaje, un gestor de recursos novato desconoce su potencial y por lo tanto aceptará todos los contratos adjudicados para la provisión de recursos sin discriminar las tareas, incurriendo en elevados costos adicionales que reducirán considerablemente beneficios y credibilidad, debido a los incumplimientos. Esta situación se irá revirtiendo a través de sucesivas interacciones que le permitirán al servidor reconocer a que tareas es capaz de proveerle recursos.

El mecanismo de emparejamiento está basado en el algoritmo de *Q-learning*.

Cada gestor de fines tiene una lista de valores Q asociado a un agente del otro tipo y a un recurso en particular.

Por cada emparejamiento que finaliza, los agentes involucrados reciben la recompensa correspondiente y se actualiza su valor de Q. Para elegir un socio los agentes utilizan una política ε-greedy. Esta política define dos acciones básicas para un agente: puede explorar, con una probabilidad de ε, o explotar, con una probabilidad de 1-ε. En el caso que el agente decida explotar, utilizará su conocimiento previo para tomar las decisiones. La elección de su socio se basa en el valor de Q que tenga para dicho socio y el recurso que esté solicitando/ofreciendo y se le da mayor prioridad al que tenga mejor valor de Q. En el caso de la explotación, la decisión que toma el agente es aleatoria, es decir, elige a cualquier agente sin preferencias.

La figura 2 muestra un pseudicódigo del algoritmo.

Al comenzar la simulación todos los valores de *Q* se inicializan en 0.

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).

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

$r_{t+1}$ es la recompensa obtenida de dicha asociación y $B_{t+1}$ es el factor de credibilidad que tiene el gestor de medios con quien estableció la relación cliente-servidor. Este factor evalúa cuán bien el servidor se desempeñó. Para los restantes gestores de medios se mantiene el valor de Q.

El factor de credibilidad se establece inicialmente en 100 para todos los gestores servidores, 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).

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

**Para todos** los gestores:
Inicializar los valores de Q en 0 y las recompensas de forma aleatoria

**Repetir** T veces
**Para todas** las tareas que comienzan esa iteración
El GF decide si explora o explota.
**Si** explora:
Elige un GM para cada tarea al azar.
**Sino**, si explota:
Para cada recurso, elige a aquellos GM cuyo Q sea mayor al Q promedio (es decir, elige a varios gestores de fines por recurso).
Los GM analizan las solicitudes recibidas y deciden a cuál satisfacer utilizando e-greedy.
Los GF eligen al mejor de los que aceptaron satisfacerlo
**Para todas** las tareas que finalizan:
Todos los gestores reciben sus recompensas
Actualizan sus valores de Q.

**Figura 2.** Algoritmo utilizado por los gestores para la toma de decisiones.

Se considera proveedor capaz a aquel que puede proveer efectivamente el recurso solicitado. De esta manera se busca que los gestores aprendan a elegir entre aquellos que son capaces de proveer los recursos requeridos y a descartar aquellos que no, asignándoles un valor de *Q* más bajo, penalizándolo de esta manera.

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. Nuevamente, al iniciar las iteraciones el gestor de medios desconoce las tareas para las cuales es capaz de proveer recursos más eficazmente.

$$Q_{t+1} = Q_t + \propto [r_{t+1} - Q_t] \tag{3}$$

$$r_t = B_t - P_t \tag{4}$$

$$P_{t+1} = \begin{cases} P_t, \ si\ es\ capaz \\ P_t + 25, \ si\ no\ es\ capaz \end{cases} \tag{5}$$

En el pseudocódigo además puede verse que en el caso de que el gestor de fines decida explotar, no elige a un solo gestor de medios sino a un conjunto de ellos, aquellos cuyo valor de Q sea mayor al valor promedio. De esta manera, el gestor comienza enviando solicitudes a todos los posibles socios (ya que todos tienen el mismo valor de Q), y a medida que va aprendiendo este conjunto se va reduciendo a los proveedores que considera más convenientes.

# 4 Resultados

Para obtener los resultados de las simulaciones del modelo, se realizó una serie de corridas con diferentes configuraciones, a fin de observar el comportamiento de los agentes. En todos los casos se utilizó $\alpha = 0.1,\ \varepsilon = 0.1$ y corridas de 1000 iteraciones. Estos tres valores fueron obtenidos de manera experimental, observando el comportamiento de los agentes con diferentes combinaciones de valores.

## 4.1 Los Gestor de Fines aprenden a seleccionar proveedor

En primer lugar, se realizó una corrida con 2 gestores de fines y 5 gestores de medios (proveedores). Se observó el porcentaje de elección de proveedores, para un gestor de fines y un recurso en particular.

De los 5 proveedores, se tiene que sólo los proveedores 1, 3 y 5 son capaces de proveer el recurso, y a su vez el 3 es el que brinda una mayor recompensa.

En la figura 3 se puede observar cómo a medida que transcurren las iteraciones el gestor de fines aprende a seleccionar al proveedor más conveniente, en este caso el 3. En primer lugar aprende a diferenciar qué gestor de medios es capaz de proveer el recurso, y por otro lado a seleccionar al más conveniente.

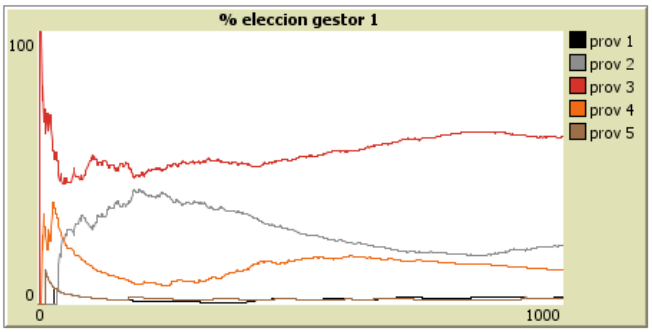


**Figura 3**. Elección del gestor de fines 1.

## 4.2 Los proveedores aprenden a aceptar relaciones beneficiosas

La segunda corrida se realizó con un escenario compuesto de 5 gestores de fines y 2 gestores de medios. En este caso, se observa el grado en que un gestor de medios acepta a un gestor de fines para proveerle un recurso. Las recompensas dadas por los gestores de fines para dicho recursos son: GF1: 121, GF2: 35, GF3: 167, GF4: 79, GF5: 98.

En la figura 4 se ve cómo el gestor de medios elige un mayor porcentaje de veces al gestor de fines 3, que es el que mayor recompensa le brinda.

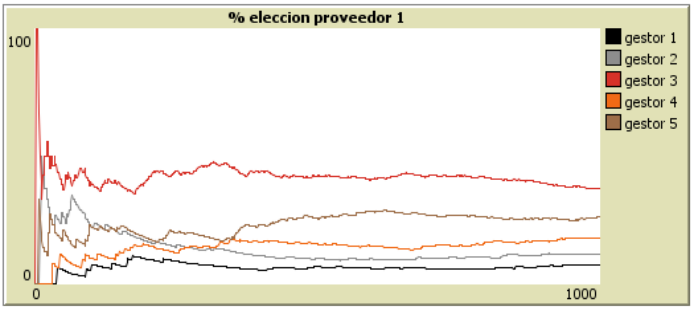


**Figura 4.** Elección del gestor de medios 1.

Como se puede ver en las figuras 3 y 4, la preferencia de un gestor de fines sobre un gestor de medios es mayor que la preferencia de un gestor de medios sobre un gestor de fines. Mientras que en la corrida 1, al cabo de 1000 iteraciones el gestor de fines seleccionó al mejor proveedor un 61% de las veces, en la corrida 2, el gestor de medios elige a su socio más conveniente un 37% de las veces.

Este comportamiento se repite en las diferentes corridas, y se debe a que el proveedor depende de la definición de tareas del gestor de fines. Es probable que en algún instante, el gestor que él prefiere no defina ninguna tarea que requiera de sus recursos, y por lo tanto deberá negociar con otros gestores, aunque obtenga menor recompensa.

## 4.3 Recompensas obtenidas

En los dos escenarios descriptos anteriormente, se observó la recompensa promedio obtenido por los gestores de cada tipo.

En el primer caso, se tenía mayor cantidad de gestores de fines que de medios, y en el segundo viceversa.

Las figuras 5 y 6 muestran que en ambos casos los gestores menos numerosos son los que consiguen en promedio una mayor recompensa. Por el contrario, cuando hay más gestores de su mismo tipo, los gestores reciben una recompensa menor. Es decir que cuanta más competencia haya entre los gestores, menor será su recompensa, mientras que al no tener competencia y una amplia demanda, es más libre para seleccionar a su socio, y por lo tanto puede hacer una mejor elección.

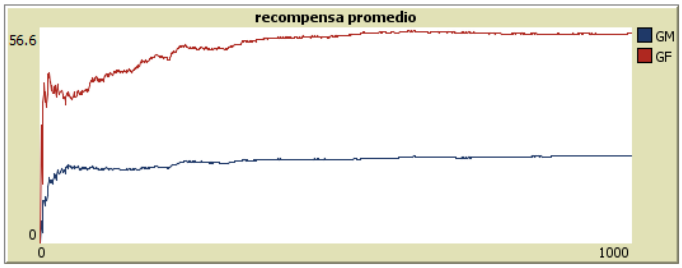


**Figura 5**. Recompensa promedio con más gestores de medios.


**Figura 6.** Recompensa promedio con más gestores de fines.

## 4.4 Algoritmo con inteligencia vs algoritmo sin inteligencia

Para poder probar la efectividad de la incorporación de aprendizaje en los agentes, se realizó un modelo similar, donde la elección de un socio se hace de manera aleatoria.

Se realizaron 20 corridas de ambos modelos. La figura 7 muestra una comparación de las recompensas obtenidas en promedio por cada tipo de gestor por corrida.

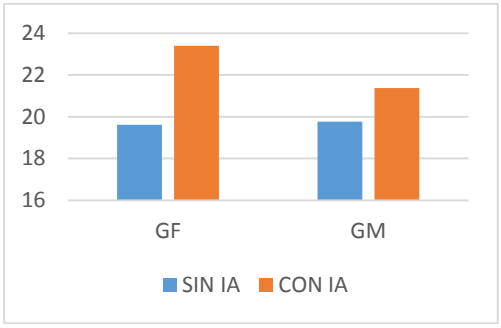


**Figura 7**. Recompensa promedio obtenida

## 5 Conclusiones

El trabajo presenta la incorporación de la capacidad de aprendizaje en los agentes gestores en el marco de compañía fractal basada en proyectos. La estructura fundamental del modelo de empresa son relaciones cliente-servidor que los gestores de proyectos establecen para el logro de un determinado fin o la provisión de un medio. La eficiencia y eficacia de los resultados obtenidos en el logro de sus metas dependerá de la estrategia de selección de socios que emplee cada gestor de proyecto. Por ello, se plantea la incorporación de aprendizaje a los agentes, de manera que puedan lograr las relaciones más beneficiosas para sus negocios.

En las simulaciones realizadas del modelo, se han podido observar diferentes comportamientos de los agentes. En primer lugar, puede verse que los agentes aprenden a seleccionar el socio que más le conviene para negociar un recurso determinado. Los agentes de ambos tipos intentan maximizar sus recompensas, para lo cual se basan tanto en el beneficio obtenido como en la confiabilidad de los agentes del otro tipo. También puede verse la mejora que representa una toma de decisiones inteligente frente a un modelo sin inteligencia, donde las decisiones se toman de manera aleatoria.

Si bien este trabajo en particular se basa en un modelo sencillo de la compañía fractal basada en proyectos, esta simplificación nos provee un útil punto de partida para estudiar la incidencia del aprendizaje en los agentes gestores de proyecto en mercados bilaterales y cómo los mismos logran conocer sus preferencias y entorno para lograr coincidencias entre gestores clientes y servidores. 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.

## Referencias

1. Canavesio, M, Martinez, E. Enterprise modeling of a Project-oriented fractal company for SMEs networking. Computer in Industry Nro 58. (2007) 794-813.
2. Sotomayor, M. Implementation in the many-to-many matching market. Games and economic behaviour. Nro 46. (2004) 199-212.
3. Quaglia, C., Canavesio, M., Martinez, E. Agentes inteligentes aprenden a establecer relaciones entre socios en mercados bilaterales. Revista electrónica Argenitna-Brasil de Tecnologicas de la Información y la Comunicación REABTIC. ISSN 2446-7634. Vol 1. Nro 2. (2014).
4. Warnecke, H.J. The fractal campany. A revolution incorporate culture. Springer-Varlag. Berlin. (1993)
5. Sarne, D., Kraus, S, Managing parallel inquiries in agents` two-sided search. Artificial intelligent Vol 172 (4-5) (2008) Pp 541-569.
6. Rochet,J, Tirole, J. Tying in two-sided markets and the honor all cards rule. International journal of industrial organization Vol 26 Nro 6 (2008) Pp 1333-1347.
7. Kumar,R., Lifshits,Y., Tomkins,A. Evolution of two-sided markets. ACM Proceeding of the third ACM international Conference on web search and dataming. ISBN 978-1-60558889-6 (2010) PP. 311-320.
8. Chen, J., Song, K., Two-sided matching in the loan market. International journal of industrial organization Nro 33 (2013) Pp. 145-152.
9. Sutton, R. Barto,A Reinforcement learning. An introduction MIT Press. (1998)
10. Rochet, J, Tirole, J.. Two-sided markets: a progress report. The RAND journal of economics Vol 37 (3) (2006) Pp 645-667.
11. Das, S., Kamenica,E., Proceeding 19th Intenational Joint Conference on Artificial Intelligence IJCAI'05. (2005) Pp. 947-952.
12. Das, S., Dealers, Insiders, and Bandits: learning and its effects on market outcomes. PhD Thesis. Massachusetts Institute of Technology. (2006).