Coordinador de sistemas multiagente en entornos de toma de decisiones secuenciales

por

Patricia Paredes Hernández

Tesis sometida como requisito parcial para obtener el grado de MAESTRO EN CIENCIAS EN LA ESPECIALIDAD DE CIENCIAS COMPUTACIONALES en el Instituto Nacional de Astrofísica, Óptica y Electrónica

Supervisada por:

Dr. José Enrique Muñoz de Cote Flores
Dr. Eduardo Morales Manzanares

Tonanzintla, Puebla.

©INAOE 2014

Derechos reservados

El autor otorga al INAOE el permiso de reproducción y distribuir copias en su totalidad o en partes de esta tesis

Agradecimientos

Este trabajo no habría sido posible sin la colaboración de mis asesores el Dr. Enrique Muñoz de Cote y el Dr. Eduardo Morales Manzanares, bajo cuya supervisión escogí este tema y comencé la tesis. Gracias por guiarme en el inmenso mundo de la ciencia y por brindarme su apoyo y disposición en cada etapa de esta investigación.

Un agradecimiento a mis revisores Dr. Felipe Orihuela Espina, Dr. Manuel Montes y Gómez y Dr. Leopoldo Altamirano Robles, por su comprensión y sus consejos que fueron fundamentales para hacer posible este trabajo.

A quienes creyeron en mí, mis padres Elfego Parede Funez y Ofelia Hernández Flores y a mis hermanos Mauricio, Leticia, Sandra y Carlos. Gracias por su apoyo invaluable a lo largo de esta aventura.

Al pueblo de México que, mediante el Consejo Nacional de Ciencia y Tecnología, me apoyó con una beca para realizar los estudios de maestría.

Al Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE) por facilitarme las herramientas necesarias durante mi estancia como estudiante de maestría.

A mi esposo Hugo Fernández Hernández y a nuestra hija Isabel Fernández Paredes por su amor y estímulo constante que me inspirarán a emprender cualquier reto, a quienes de manera especial dedico este trabajo.

Resumen

La necesidad actual de resolver problemas complejos que involucran diversas tareas ha definido a los Sistemas Multiagente (SMA) como uno de los temas de investigación más relevantes en los últimos años. Un SMA está compuesto por un número determinado de agentes, que interactúan para resolver problemas, que están más allá de las capacidades individuales de cada uno de ellos. El éxito de un SMA en la solución de un problema depende esencialmente de la coordinación entre los agentes para lograr una meta común. La mayoría de las propuestas de coordinación en SMA asumen que los gentes pueden comunicarse entre ellos, que pueden entenderse (tienen un protocolo de comunicación común) y en base a la comunicación logran coordinarse. En escenarios donde no existe comunicación directa entre los agentes, caso común de los SMA abiertos, es necesario idear una forma de coordinación, donde la comunicación directa entre los agentes no sea necesaria. Una forma de atacar este problema es mediante un coordinador central, sin embargo contrarresta la autonomía de los agentes en la toma de decisiones. En este trabajo, se propone un sistema de coordinación central que propicie la toma de decisiones distribuidas, lo cual permite que cada agente mantenga su autonomía en la toma de decisiones. Nuestra propuesta se basa en un mecanismo de incentivos que le permite motivar a los agentes a cambiar su política individual por una política conjunta que maximice el valor esperado del sistema. Además, se diseñan tales incentivos con el fin de realizar una inversión mínima. El sistema se evalúa en dos escenarios simulados bajo diferentes condiciones en donde se prueba la efectividad de la propuesta.

Abstract

The current need to solve complex problems involving several tasks, has boot the Multiagent Systems area (MAS) as one of the most important research topics in recent years. A MAS is composed of a number of agents that interact to solve problems that are beyond their individual capabilities. The success of a MAS in solving a problem essentially depends on the coordination between agents to achieve a common goal. Most of the coordination schemes proposed in MAS assume that agents can communicate between them, i.e . they can understand (they have a common communication protocol) and based on this communication they can achieve coordinated. In scenarios where there is no direct communication between agents, which is common in open MAS, it is necessary to devise alternative forms of coordination, where the direct communication between the agents is not necessary. One way to address this problem is through a central coordinator, however this counteracts the autonomy of agents in decision making. In this research, we propose a novel central coordination system that fosters distributed decision making and allows each agent to maintain its autonomy in decision-making. Our proposal is based on an incentive mechanism that motivates the agents to change their individual policy to a joint policy that maximizes the expected value of the system. In addition, we introduce a novel scheme to minimize the ammount of incentives. The system is evaluated in two simulated scenarios under different conditions where the effectiveness of the proposal is shown.

Índice general

Agradecimientos I

Resumen III

Abstract IV

Índice de figuras VIII

Índice de tablas IX

Índice de algoritmos X

1. Introducción 1

1.1. Definición del problema 3
1.2. Objetivo general 4
1.3. Objetivos específicos 4
1.4. Contribuciones 5
1.5. Organización de la tesis 5

2. Antecedentes 7

2.1. Agente 7
2.2. Sistema Multiagente 8
2.2.1. Sistemas multiagente abiertos 9
2.3. Teoría de decisión 10
2.4. Proceso de decisión de Markov 11
2.4.1. Comportamiento óptimo en decisiones secuenciales 12
2.5. Métodos de solución de MDPs 13

2.5.1. Programación dinámica 13
2.5.2. Monte Carlo 16
2.5.3. Aprendizaje por refuerzo 17
2.6. Teoría de Juegos 20
2.6.1. Diseño de mecanismos 20

3. Trabajo relacionado 23
3.1. Coordinación 23
3.1.1. Dependencias entre actividades de los agentes 24
3.1.2. Razones para coordinar 26
3.2. Enfoques de coordinación en un SMA 26
3.2.1. Coordinación multiagente en un enfoque descentralizado . . . 27
3.2.2. Técnicas de Coordinación 27
3.2.3. Coordinación multiagente en un enfoque centralizado 30

4. Modelo propuesto 33
4.1. Sistema de Coordinación propuesto 34
4.1.1. Sistema multiagente 34
4.1.2. Coordinador 36
4.1.3. Mecanismo de incentivos 37

5. Experimentos y resultados 41
5.1. Caso de estudio 41
5.2. Escenario experimental 42
5.3. Descripción de los experimentos 45
5.4. Experimento 1: Comportamiento de los agentes sin coordinador . . . 46
5.4.1. Resultados 46
5.4.2. Discusión de resultados 50
5.5. Experimento 2: Agentes deciden reportar su estado falso o verdadero. 51
5.5.1. Resultados 51
5.5.2. Discusión de resultados 54
5.6. Experimento 3: Agentes deciden reportar estado privado falso o verdadero y en obedecer o no al coordinador 55
5.6.1. Resultados 55

5.6.2. Discusión de resultados 58
5.7. Experimento 4: Evaluación del mecanismo de incentivos 58
5.7.1. Resultados 59
5.7.2. Discusión de resultados 61
5.8. Experimento 5: Ajuste del presupuesto del coordinador 62
5.8.1. Resultados 63
5.8.2. Discusión de resultados 65
5.9. Experimento 6: Redistribución de utilidades 66
5.9.1. Resultados 67
5.9.2. Discusión de resultados 69

6. Conclusiones y trabajo futuro **72**
6.1. Conclusiones 72
6.2. Trabajo futuro 74

Bibliografía **77**

Índice de figuras

2.1. Modelo clásico de interacción agente-entorno. 18

4.1. Sistema de coordinación propuesto 35

5.1. Escenarios experimentales. 43
5.2. Visibilidad de los agentes. 44
5.3. Política conjunta óptima 45
5.4. Escenario 1: agentes sin coordinador, penalización de -5 47
5.5. Escenario 1: agentes sin coordinador, penalización de -10 47
5.6. Escenario 2: agentes sin coordinador, penalización de -5 49
5.7. Escenario 2: agentes sin coordinador, penalización de -10 49
5.8. Escenario 1: agentes reportan información falsa 52
5.9. Escenario 2: agentes reportan información falsa 52
5.10. Escenario 1: agentes siguiendo siguiendo su propia política 56
5.11. Escenario 2: agentes siguiendo siguiendo su propia política 56
5.12. Escenario 1: agentes siguiendo política conjunta 60
5.13. Escenario 2: agentes siguiendo política conjunta 60

Índice de tablas

5.1. Definición de los MDPs de los agentes. 42
5.2. Políticas individuales para el escenario 1 48
5.3. Políticas individuales para el escenario 2 50
5.4. Políticas resultantes cuando reportan información falsa 53
5.5. Políticas resultantes cuando ignoran al coordinador 57
5.6. Política conjunta, al implementar el mecanismo de incentivos 61
5.7. Ajuste de presupuesto para escenario 1 64
5.8. Ajuste de presupuesto para escenario 2 64
5.9. Distribución de utilidades en escenario 1 68
5.10. Distribución de utilidades en escenario 2 68

Índice de algoritmos

1. Iterativo de evaluación de política 14
2. Iteración de política 15
3. Iteración de valor 15
4. Algoritmo para estimar V^π 16
5. Algoritmo de Monte Carlo 17
6. Algoritmo Q-Learning 19
7. Mecanismo de incentivos propuesto 38

Capítulo 1

Introducción

En la actualidad existen diversos dominios en los que la de división y distribución de tareas es fundamental para la solución de problemas. Por tal motivo, la Inteligencia Artificial Distribuida (IAD) ha sido una de las áreas de investigación más importantes en los últimos años.

La IAD es un campo de la Inteligencia Artificial dedicado al estudio de las técnicas y métodos necesarios para la coordinación y distribución del conocimiento y las acciones en un entorno con múltiples agentes.

En la comunidad investigadora no existe acuerdo sobre el concepto de agente, existiendo varias definiciones sobre el mismo. La definición de M. Wooldridge en [16] es la que tiene una mayor aceptación, definiendo un agente como:

Una entidad física o abstracta que puede percibir su ambiente, es capaz de evaluar tales percepciones y tomar decisiones por medio de mecanismos de razonamiento simples o complejos, comunicarse con otros agentes para obtener información y actuar sobre el medio en el que se desenvuelven a través de ejecutores.

De esta definición podemos destacar que un agente debe ser capaz de interactuar, con su entorno, con el fin de conseguir de manera autónoma ciertos objetivos para los cuales ha sido diseñado. Teniendo en cuenta estos aspectos, un Sistema Multiagente (SMA) es definido por K. P. Sycara en [12] como:

Un conjunto de agentes que interaccionan para resolver problemas, que están más allá de las capacidades individuales de cada uno de ellos.

El éxito de un SMA en la solución de un problema depende esencialmente de la coordinación entre los agentes. La coordinación es el proceso por el cual un agente razona acerca de sus acciones locales y las acciones (anticipadas) de los otros. Se busca asegurar una comunicación que permita a los agentes llevar a cabo sus acciones y comportamientos de forma coherente.

Una coordinación efectiva es esencial si se quiere que diferentes agentes autónomos logren alcanzar un objetivo en común. Dicha coordinación requiere gestionar las diferentes formas de dependencia que se producen cuando los agentes: 1) tienen objetivos en común 2) comparten un entorno, ó 3) hay recursos compartidos [21]. El objetivo de la coordinación es lograr una integración y un ajuste armonioso de los esfuerzos de trabajo individual en beneficio de una meta común.

Dentro del área de SMA, la coordinación distribuida entre los agentes se puede dar mediante: la cooperación o negociación [13]. La cooperación es la coordinación entre agentes no antagonistas, que resuelven un problema común colaborando mutuamente y la negociación es la coordinación entre agentes auto-interesados o simplemente competitivos que a pesar de no tener un objetivo en común, la comunicación entre ellos les permite generar acuerdos. Lógicamente es la comunicación la que posibilita a los agentes coordinar sus acciones y comportamiento, y obtener como resultado sistemas coherentes, es decir sistemas que se comportan bien como una unidad.

Al principio, la mayor parte de la investigación sobre estos sistemas se centraba en los denominados sistemas cerrados, donde se tiene un control sobre los agentes, y la infraestructura que forman dichos sistemas [14]. Por tanto, se puede asegurar que el comportamiento de los agentes será correcto y acorde al objetivo global de estos sistemas. Sin embargo, no podemos esperar que esta suposición se cumpla cuando tratamos de pasar estas ideas a SMA abiertos [3].

En un SMA abierto los agentes son construidos por distintos diseñadores y/o representan entidades diferentes del entorno, actúan de forma racional persiguien-

do un objetivo individual y además puede no existir comunicación directa entre ellos.

Debido a que la comunicación entre los agentes es clave para la coordinación coherente, en un SMA abierto donde no existe comunicación directa entre los agentes la coordinación a nivel distribuido es inadecuada, ya que cada agente intenta cumplir su objetivo individual entorpeciendo, sin darse cuenta, el de los demás dando como resultado que el problema global no sea solucionado o la solución no sea la esperada.

Una opción para este tipo de sistemas es coordinar a los agentes de manera centralizada. Es decir, mediante un planeador central, el cual a partir de información recibida de los agentes, define lineas de acción en beneficio de cada uno de ellos.

1.1. Definición del problema

La coordinación centralizada en un SMA, contraviene con el concepto de agente autónomo, el cual lo identifica como una entidad capaz de tomar decisiones de manera individual.

Algunos trabajos [23], [4], han considerado representar esta autonomía en la toma de decisiones, cuando los agentes revelan información privada al planeador central, es decir consideran que cada agente puede decidir en revelar información completa (información verdadera) o retener aquella que pueda representarles alguna desventaja ante los demás (información falsa). Pero finalmente la decisión respecto a las acciones a realizar esta a cargo del planeador central.

En este trabajo se diseña e implementa un sistema de coordinación central que propicia la toma de decisiones distribuidas, lo cual permite que cada agente mantenga su autonomía en la toma de decisiones.

Para lograr este objetivo nos basamos en un mecanismo de incentivos, que permite al coordinador motivar a los agentes a: 1) decidir participar voluntariamente como parte del SMA, 2) decidir compartir su información privada y 3) cambiar su política individual por una política conjunta que maximice el valor esperado del sistema

además, se desea lograr diseñar tales incentivos de tal forma que se realice una inversión mínima.

1.2. Objetivo general

Diseñar e implementar un sistema de coordinador multiagente en un ambiente con toma de decisiones secuenciales, entre agentes con intereses propios, capaz de obtener información privada verídica de los agentes y de convencerlos a participar voluntariamente cambiando su política individual por una política conjunta que maximice el valor esperado del sistema, sin que los agentes pierdan su racionalidad individual.

1.3. Objetivos específicos

- Diseñar e implementar un ambiente multiagente simulado, donde exista toma de decisiones secuenciales.

- Diseñar e implementar un coordinador capaz de fusionar los estados de cada agente del sistema y emitir una acción conjunta en cada periodo de decisión.

- Definir un mecanismo de incentivos que permita al coordinador convencer a los agentes de:

    1. Revelar información verídica.
    2. Seguir sus recomendaciones.

- Definir una función de pagos que le permita al coordinador distribuir la utilidad que los agentes reciben del sistema para coordinarlos, realizando una inversión monetaria mínima.

- Mostrar la efectividad del sistema en un ambiente multiagente simulado.

1.4. Contribuciones

A continuación se enuncian las principales contribuciones de este trabajo de investigación.

- Un coordinador capaz de definir una política conjunta que maximice el valor esperado del sistema, a través de la información recibida de los agentes.
- Un mecanismo de incentivos que le permita al coordinador obtener no solo información verídica de los agentes si no también, convencerlos de cambiar su política individual por una política conjunta que maximice el valor esperado del sistema, sin que los agentes pierdan su racionalidad individual.
- Una función de pagos que le permita al coordinador distribuir la utilidad que los agentes reciben del sistema para coordinarlos, realizando una inversión mínima.

1.5. Organización de la tesis

A continuación exponemos cómo queda estructurada la presente tesis:

En el capítulo 2 se presentan los conceptos básicos necesarios para entender el tema de agentes, sistemas multiagente abiertos, coordinación multiagente y diseño de mecanismos, así como la descripción de los modelos de aprendizaje básicos utilizados en esta tesis.

En el capítulo 3 se presenta una revisión del problema de coordinación en SMA, tema principal de la presente tesis. Se describen y comparan dos clasificaciones de coordinación respecto a los tipos de SMA y los principales trabajos de investigación desarrollados para dar solución a este problema. Finalmente, se define el enfoque sobre el cual se basa esta investigación y sus principales contribuciones.

En el capítulo 4 se describe el sistema de coordinación propuesto. Se presenta el diseño general del coordinador, el modelo de los agentes del SMA diseñado y el mecanismo de incentivos propuesto.

En el capítulo 5 se describe el caso de estudio sobre el cual se evalúa el sistema de coordinación propuesto. Se detallan los experimentos que fueron realizados en dos etapas: la primera con el objetivo de analizar el comportamiento de los agentes de forma individual y al coordinador propuesto sin el mecanismo de incentivos, en la segunda etapa se evaluar el coordinador con el mecanismo de incentivos propuesto y el análisis que permite diseñar los incentivos tal que, el coordinador realice una inversión mínima.

Finalmente, en el Capítulo 6 se presentan las conclusiones de la tesis, así como las principales líneas de trabajo futuro.

Capítulo 2

Antecedentes

En este capítulo se presenta una visión del conocimiento más relevante en las diversas áreas que guardan relación con el presente trabajo. Se definen los siguientes temas: agente, sistemas multiagente, coordinación, aprendizaje automático, teoría de juegos y diseño de mecanismos.

2.1. Agente

Retomando la definición de agente de M. Wooldridge en [16], expresada en la introducción, podemos destacar que un agente es una entidad física o virtual el cual desempeña alguna actividad en un ambiente determinado. El agente es capaz de percibir su entorno en forma de estados e interactúa con él mediante la toma de decisiones, cumpliendo con las siguientes características:

- Capacidad para resolver problemas no triviales: un agente inteligente sabe razonar sobre el entorno, capacidad que le permite realizar un conjunto de tareas.
- Racionalidad (limitada): Los agentes están dotados de un conjunto de objetivos y emprenden acciones para cumplirlos. Eligen sus acciones según el principio de racionalidad, es decir, prefieren la acción más prometedora para sus metas. Sin embargo, su actuación está limitada por el tiempo y los recursos.
- Autonomía (limitada): Los agentes tienen sus propias motivaciones a partir de las que generan autónomamente sus objetivos. Sin embargo, para alcanzar

gran parte de los mismos dependen de la ayuda de los demás agentes, lo cual impone un límite a su autonomía.

- Reactividad y Pro-Actividad: Los agentes perciben el entorno y responden a los cambios que ocurren en él. Además, debido a que generan sus propias metas y pueden actuar convenientemente son capaces de tomar la iniciativa.

2.2. Sistema Multiagente

Un Sistema Multiagente (SMA), es una colección de agentes que trabajan juntos para resolver problemas que están más allá de las capacidades individuales o conocimiento de cualquiera de ellos [19]. Para que un SMA cumpla su objetivo global a partir de los objetivos individuales de los agentes participantes, es necesario coordinar sus acciones de lo contrario el objetivo global no será cumplido o el resultado final no será el esperado.

La coordinación es tal vez uno de los problemas clave en Inteligencia Artificial Distribuida (IAD) y por lo tanto de los SMA. Se denomina *coordinación* al proceso por el cual un agente razona acerca de sus acciones locales y las acciones (anticipadas) de los otros, en un intento de asegurar una comunicación que permita a los agentes llevar a cabo sus acciones y comportamientos de forma coherente. Una coordinación efectiva es esencial si diferentes agentes autónomos quieren alcanzar sus objetivos dentro de un SMA. Dicha coordinación requiere gestionar las diferentes formas de dependencia que se producen cuando los agentes: 1) tienen objetivos inter-enlazados, 2) comparten un entorno común, ó 3) hay recursos compartidos [21].

Lógicamente con el fin de alcanzar dicha coordinación se han desarrollado hasta la fecha una gran diversidad de mecanismos donde, la comunicación es la que posibilita a los agentes coordinar sus acciones y comportamiento. En estos mecanismos los protocolos de interacción gobiernan el intercambio o la comunicación de una serie de mensajes entre agentes (una conversación). En aquellos casos en que los agentes son auto-interesados o tienen objetivos conflictivos -negociación- el principal objetivo de los protocolos es maximizar los pagos o utilidades de los agentes. En aquellos casos

en que los agentes tienen similares objetivos o problemas comunes –cooperación-, el objetivo de los protocolos es mantener un comportamiento de los agentes globalmente coherente sin violar su autonomía.

Dichos mecanismos de coordinación han tenido éxito en los denominados SMA cerrados donde, se tiene un control, tanto sobre los agentes como sobre la infraestructura que forman dichos sistemas [14]. Por tanto, se puede asegurar que mediante un protocolo de comunicación adecuado, el comportamiento de los agentes será correcto y acorde al objetivo global de estos sistemas. Sin embargo, no podemos esperar que esta suposición se cumpla cuando hablamos de SMA abiertos [3], ya que no se tiene control de los agentes ni de la infraestructura del sistema.

2.2.1. Sistemas multiagente abiertos

En los sistemas multiagente abiertos, los agentes pueden ser construidos por distintos diseñadores y/o representar intereses de entidades diferentes del mundo. Algunas características distintivas de un SMA abierto son:

- No existe un objetivo global.
- Los agentes son autónomos y distribuidos.
- Cada agente tiene información incompleta o capacidades limitadas, por los que cada agente tiene un punto de vista limitado.
- No existe un diseño centralizado.
- No existe un control global del sistema.
- No existe comunicación directa entre los agentes, por lo que su información es privada.

Debido a estas características, en un SMA abierto, los agentes presentan un comportamiento individualista es decir, basados en su racionalidad cada uno elegirán la mejor estrategia para él mismo sin importar la de los demás. En algunas investigaciones enfocadas en este tipo de sistemas, a los agentes que presentan este

comportamiento les denominan *agentes egoístas*. En este trabajo les denominamos *agentes estratégicos*.

Como se ha mencionado, los agentes interactúan con su entorno mediante la toma de decisiones, para que un agente cumpla con la característica de racionalidad debe evaluar sus decisiones antes de ser aplicadas en el entorno para lograrlo, los agentes modelan este proceso mediante la teoría de decisión.

2.3. Teoría de decisión

Decisión es una elección entre dos o más líneas de acción diferentes. El objeto de la teoría de decisión es racionalizar dicha elección. Para ello hay que proceder de modo sistemático, evaluando todas las posibilidades de elección como las consecuencias que puedan derivarse de cada opción.

Los agentes basados en teoría de decisión son capaces de tomar decisiones racionales motivadas por maximizar una utilidad esperada. Dentro de la teoría de decisión, las decisiones se clasifican en estáticas o secuenciales.

- *Decisiones estáticas:* son aquellos problemas de decisión en los que sólo se adopta una decisión y no se analizan las posibles alternativas/decisiones posteriores a una alternativa/decisión previa.
- *Decisiones secuenciales:* problema de decisión en el que se considera una secuencia de decisiones, es decir, decisiones posteriores dependientes de una decisión inicial por lo tanto, la utilidad para el agente depende de la secuencia de acciones.

En esta tesis nos enfocamos a SMA abiertos donde los agentes modelan problemas de decisiones secuenciales donde, el comportamiento del agente debe de ser tal que escoja acciones que tiendan a incrementar a largo plazo la suma de las recompensas totales.

En general en este tipo de problemas, el entorno es *no-determinístico* (tomar la misma acción en el mismo estado puede dar resultados diferentes). Sin embargo, se

asume que el entorno es *determinístico* o estacionario (esto es, las probabilidades de cambio de estado no cambian o lo hacen muy lentamente).

Otro aspecto importante es el balance entre exploración y explotación. Para obtener buena ganancia un agente prefiere seguir ciertas acciones pero, para saber cuáles se tiene que hacer cierta exploración. Muchas veces depende de cuánto tiempo se espera que el agente interactúe con el entorno. La caracterización de esta problemática está dada por procesos de decisión de Markov o MDP.

2.4. Proceso de decisión de Markov

Un proceso de decisión de Markov (Markov Desicion Process MDP) es un modelo probabilista para toma de decisiones secuenciales en donde el sistema evoluciona en el tiempo y es controlado por un agente.

La dinámica del sistema está determinada por una función de transición de probabilidad que mapea estados y acciones a otros estados. Formalmente, un MDP es una tupla $M =< S, A, \Phi, R >$. Los elementos de un MDP son:

- Un conjunto finito de estados $S(1, ..., n)$.
- Un conjunto finito de acciones A, que pueden depender de cada estado.
- Φ es una función de transición de estados dada como una distribución de probabilidad $\Phi : S \times A \times S \rightarrow [0, 1]$. La probabilidad de alcanzar el estado $s' \in S$ al realizar la acción $a \in A$ en el estado $s \in S$, que se puede denotar como $\Phi(a, s, s')$.
- Función de recompensa R: define la meta. Mapea cada estado–acción a un número (recompensa) $R : S \times A \rightarrow \mathbb{R}$, indicando lo deseable del estado.

Un MDP define una política π ómo el comportamiento del sistema en cierto tiempo. π Es un mapeo de los estados a las acciones. Una politica optima π^* es aquella que permite obtener la función de valor V misma que indica lo que es bueno a largo plazo. V Es la recompensa total que un agente puede esperar acumular

empezando en ese estado (predicciones de recompensas). Se busca tomar acciones que den los valores más altos, no la recompensa mayor.

2.4.1. Comportamiento óptimo en decisiones secuenciales

El objetivo del agente es maximizar la recompensa a largo plazo. Las recompensas obtenidas a partir de un tiempo t se denotan como: $r_{(t+1)}, r_{(t+2)}, r_{(t+3)}....$ Entonces, lo que se desea maximizar es $R(T) = r_{(t+1)} + r_{(t+2)} + r_{(t+3)} + \cdots + r_T$. Si se tiene un punto terminal se llaman tareas *episódicas*. Si no se tiene se llaman tareas *continuas*. Este último caso representa un problema, ya que no se puede hacer el cálculo cuando T no tiene límite.

Una alternativa para tratar este problema es utilizando una razón de descuento, la cual hace que la aportación de las recompensas más lejanas sean más pequeñas. Los tres modelos básicos de comportamiento óptimo son:

1. Horizonte finito. El agente trata de maximizar su recompensa en los próximos h pasos, sin tomar en cuenta lo que ocurra después.

$$max(\, E(\sum_{t=0}^{h} r_t\,)\,)$$

2. Horizonte infinito. Se utiliza un *factor de descuento* γ, $(0 < \gamma < 1)$, para reducir las recompensas, con el fin de evitar una suma infinita.

$$max(\, E(\sum_{t=0}^{\infty} \gamma^t r_t\,)\,)$$

3. Recompensa promedio. Optimizar a largo plazo las recompensas promedio.

$$lim_{h\to\infty} E(\,\frac{1}{h}\sum_{t=0}^{h} r_t\,)$$

2.5. Métodos de solución de MDPs

Existen tres formas principales de resolver MDPs: (1) usando métodos de programación dinámica, (2) usando métodos de Monte Carlo, y (3) usando aprendizaje por refuerzo.

2.5.1. Programación dinámica

Si se conoce el modelo del ambiente, es decir, las probabilidades de transición $\Phi(s, a, s')$ y los valores de recompensa $R(s, a, s')$, las ecuaciones de optimalidad de Bellman nos presentan un sistema de $\mid S \mid$ ecuaciones y $\mid S \mid$ incógnitas [20].

Se considera primero como calcular la función de valor V^π dada una política arbitraria π.

$$\begin{aligned} V^\pi(s) &= E_\pi\{R(t)|s_t = s\} \\ &= E_\pi\{r_{r+1} + \gamma r_{r+2} + \gamma^2 r_{r+2} + ...|s_t = s\} \\ &= E_\pi\{r_{r+1} + \gamma V^\pi(s_{t+1})|s_t = s\} \\ &= \sum_a \pi(s, a) \sum_{s'} \Phi(s, a, s')[R(s, a, s') + \gamma V^\pi(s')] \end{aligned}$$

Donde $\pi(s, a)$ es la probabilidad de tomar la acción a en el estado s bajo la política π. Por lo tanto, se pueden hacer aproximaciones sucesivas, evaluando $V_{k+1}(s)$ en términos de$V_k(s)$.

$$V_{k+1}(s) = \sum_a \pi(s, a) \sum_{s'} \Phi(s, a, s')[R(s, a, s') + \gamma V_k(s')]$$

Esto permite definir un algoritmo llamado algoritmo de evaluación iterativa de políticas descrito en el Algoritmo 1.

Una de las razones para evaluar la función de valor de una política es tratar de encontrar mejores políticas. Dada una función de valor para una política dada, podemos probar una acción $a \neq \pi(s)$ y ver si su $V(s)$ es mejor o peor que el $V^\pi(s)$.

Algoritmo 1 Iterativo de evaluación de política

Inicializar la función $V(s) = 0$ para toda $s \in S$
Repite
$\Delta \leftarrow 0$
Para cada $s \in S$
$\nu \leftarrow V(s)$
$V(s) \leftarrow \sum_a \pi(s,a) \sum_{s'} \Phi(s,a,s')[R(s,a,s') + \gamma V^{\pi}(s')]$
$\Delta \leftarrow max(\delta, |\nu - V(s)|)$
Hasta que $\Delta < \theta$ (número positivo pequeño)
Regresar $V \approx V^{\pi}$

En lugar de hacer un cambio en un estado y ver el resultado, se pueden considerar cambios en todos los estados mediante todas las acciones de cada estado, seleccionando aquella que parezca mejor de acuerdo a una política *greedy* que definiremos más adelante.

Entonces, se puede calcular una nueva política $\pi'(s) = argmax_a Q^{\pi}(s,a)$ y continuar hasta que no mejoremos. Esto sugiere, partir de una política (π_0) y calcular la función de valor (V^{π_0}), con la cual encontrar una mejor política (π_1) y así sucesivamente hasta converger a π^* y V^*. A este procedimiento se llama iteración de políticas y se describe en el algoritmo 2.

Uno de los problemas de iteración de políticas es que cada iteración involucra evaluación de políticas que requiere recorrer todos los estados varias veces. Sin embargo, el paso de evaluación de política lo podemos truncar de varias formas, sin perder la garantía de convergencia. Una de ellas es detenerla después de recorrer una sola vez todos los estados. A esta forma se le llama iteración de valor (*value iteration*). En particular se puede escribir combinando la mejora en la política y la evaluación de la política truncada como sigue:

$$V_{k+1}(s) = max_a \sum_{s'} \Phi(s,a,s')[R(s,a,s') + \gamma V_k(s')]$$

Es como expresar la ecuación de Bellman en una regla de actualización muy parecido a la regla de evaluación de políticas, solo que se evalúa el máximo sobre todas las acciones (ver algoritmo 3).

Algoritmo 2 Iteración de política

1: Inicializar las funciones $V(s) \in R$ y$\pi(s) \in A(s)$ arbitrariamente $\forall s \in S$
2: Evaluación de política:
Repite
$\Delta \leftarrow 0$
Para cada $s \in S$
$\nu \leftarrow V(s)$
$V(s) \leftarrow \sum_{s'} \Phi(s, \pi(s), s')[R(s, \pi(s), s') + \gamma V(s')]$
$\Delta \leftarrow max(\delta, |\nu - V(s)|)$
Hasta que $\Delta < \theta$ (número positivo pequeño)
3: Mejora de política:
$pol - estable \leftarrow true$
Para cada $s \in S$:
$b \leftarrow \pi(s)$
$\pi(s) \leftarrow argmax_a \sum_{s'} \Phi(s, a, s')[R(s, a, s') + \gamma V(s')]$
Si $b \neq \pi$, **entonces** $pol - estable \leftarrow false$
Si $pol - estable$, **entonces** stop, **si no** ir a 2.

Algoritmo 3 Iteración de valor

Inicializar la función $V(s) = 0$ para toda $s \in S$
Repite
$\Delta \leftarrow 0$
Para cada $s \in S$
$\nu \leftarrow V(s)$
$V(s) \leftarrow max_a \sum_{s'} \Phi(s, a, s')[R(s, a, s') + \gamma V^*(s')]$
$\Delta \leftarrow max(\delta, |\nu - V(s)|)$
Hasta que $\Delta < \theta$ (número positivo pequeño)
Regresa una política determinísta tal que:
$\pi(s) = argmax_a \sum_{s'} \Phi(s, a, s')[R(s, a, s') + \gamma V^*(s')]$

Para espacios muy grandes, el ver todos los estados puede ser computacionalmente muy caro. Una opción es hacer estas actualizaciones al momento de estar explorando el espacio, y por lo tanto determinando sobre qué estados se hacen las actualizaciones. El hacer estimaciones en base a otras estimaciones se conoce también como *bootstrapping*.

2.5.2. Monte Carlo

Los métodos de Monte Carlo, solo requieren de experiencia y la actualización se hace por episodio más que por cada paso. Un episodio es una serie de pasos (toma de decisiones) requeridos para llegar de un estado inicial a un estado final.

El valor de un estado es la recompensa esperada que se puede obtener a partir de ese estado.

Para estimar V^π y Q^π se toman estadísticas haciendo un promedio de las recompensas obtenidas. Q^π es una función que define probabilidades de transición de un agente mediante pares estado-acción, basada en la exploración del agente. La estimación de V^π se describe en el algoritmo 4.

Algoritmo 4 Algoritmo para estimar V^π

Repite
  Genera un episodio usando π
  Para cada estado s en ese episodio:
    $R \leftarrow$ recompensa después de la primera ocurrencia de s
    Añade R a $recomp(s)$
    $V(s) \leftarrow$ promedio$(recomp(s))$

Con Monte Carlo se puede alternar entre evaluación y mejoras en base a cada episodio. La idea es que después de cada episodio las recompensas observadas se usan para evaluar la política, por lo que mejora para todos los estados visitados en el episodio. Esto se describe en el algoritmo 5.

Existen dos formas para asegurar que todas las acciones pueden ser seleccionadas indefinidamente:

- Los algoritmos *on-policy*: Estiman el valor de la política mientras la usan para el control. Se trata de mejorar la política que se usa para tomar decisiones.
- Los algoritmos *off-policy*: Usan la política y el control en forma separada. La estimación de la política puede ser por ejemplo greedy y la política de com-

Algoritmo 5 Algoritmo de Monte Carlo

Repite
  Genera un episodio usando π con exploración
  Para cada par s, a en ese episodio:
    $R \leftarrow$ recompensa después de la primera ocurrencia de s, a
    Añade R a $recomp(s, a)$
    $Q(s, a) \leftarrow$ promedio$(recomp(s, a))$
  Para cada s en episodio:
    $\pi(s) \leftarrow argmax_a Q(s, a)$

portamiento puede ser $\epsilon - greedy$. Es decir, la política de comportamiento está separada de la política que se quiere mejorar.

Ejemplos de políticas de selección de acciones son:

- $\epsilon - greedy$: en donde la mayor parte del tiempo se selecciona la acción que da el mayor valor estimado, pero con probabilidad ϵ se selecciona una acción aleatoriamente.

- *softmax*: en donde la probabilidad de selección de cada acción depende de su valor estimado. La más común sigue una distribución de Boltzmann o de Gibbs, y selecciona una acción con la siguiente probabilidad:

$$\frac{e^{Q_t(a)/\tau}}{\sum_{b=1}^{n} e^{Q_t(b)/\tau}}$$

  donde τ es un parámetro positivo.

2.5.3. Aprendizaje por refuerzo

El Aprendizaje por Refuerzo (Reinforcement Learning - RL) es el problema al que se enfrenta un agente al interactuar con su entorno, con el objetivo de aprender un comportamiento a base de interacciones de prueba y error, en un entorno desconocido. Es decir, cuando el agente no conoce el modelo del ambiente (las probabilidades de transición $\Phi(s, a, s')$ y los valores de recompensa $R(s, a, s')$)

En el modelo estándar de RL el agente es un sistema de toma de decisiones ligado al ambiente mediante percepción y acción, como lo muestra la figura 2.1, que está aprendiendo a actuar a fin de maximizar su utilidad.

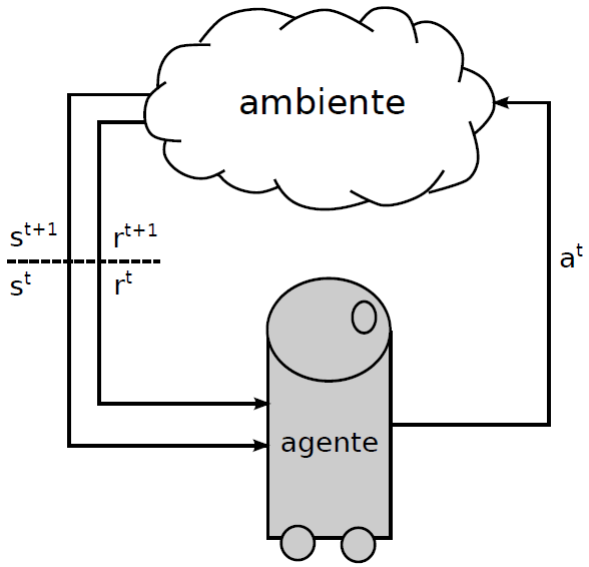

Figura 2.1: *Modelo clásico de interacción agente-entorno. El agente observa el entorno e interactúa con él realizando una acción. El ambiente cambia, o transita de estado, y retro-alimenta al agente. El agente desarrolla experiencia a partir de esta interacción y el proceso se repite.*

En cada paso o etapa de decisión, el agente identifica el estado actual, s_t, del ambiente; selecciona una acción, a_t, la cual puede modificar el estado a s_{t+1}; recibe una retroalimentación o recompensa, $r \in R$, dada por dicha transición. El comportamiento del agente debe de ser tal que escoga acciones que tiendan a incrementar a largo plazo la suma de las recompensas totales.

El modelo de aprendizaje por refuerzo para solucionar MDPs es también conocido como métodos de Diferencias Temporales (*Temporal Difference TD*). Estos métodos combinan las ventajas de los dos anteriores: permite hacer *bootstrapping* (como programación dinámica) y no requiere tener un modelo del ambiente (como Monte Carlo).

En TD usan el error o diferencia entre predicciones sucesivas (en lugar del error entre la predicción y la salida final) aprendiendo al existir cambios entre predicciones sucesivas. Sus principales ventajas son:

- Incrementales y por lo tanto fáciles de computar.
- Convergen más rápido con mejores predicciones.

Un desarrollo de gran importancia en el aprendizaje por refuerzo, es el algoritmo de Q-learning [10], [11], descrito en el Algoritmo 6.

Algoritmo 6 Algoritmo Q-Learning

Inicializar $Q(s, a)$ arbitrariamente
Repite (para cada episodio):
  Inicializa s
  Repite (para cada paso del episodio):
    Selecciona una a de s usando la politica dada por Q (e.g., $\epsilon - greedy$)
    Realiza acción a, observa r, s'
    $Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma max'_a Q(s', a') - Q(s, a)]$
    $s \leftarrow s'$;
  hasta que s sea terminal

Q-learning es un algoritmo off-policy. Es decir, no realiza la actualización de los pares de estado y acción con exactamente la misma distribución con la que los estados habrían sido encontrados siguiendo la política estimada. Q-learning garantiza converger a Q^* con una probabilidad de 1 [10], [24].

Hasta el momento, nos hemos referido a la forma en que un agente modela su entorno para resolver problemas de toma de decisiones secuenciales. Donde el comportamiento de un agente esta guiado por escoger una política π^*, que maximice su utilidad a través del tiempo.

Esta suposición ha inspirado a los investigadores para resolver el problema de la coordinación en sistemas multiagente bajo modelos de económicos y de ciencias sociales. Una de las áreas que se ha destacado en los últimos años por modelar la interacción de los agentes en cualquier sistema multiagente es la teoría de juegos.

2.6. Teoría de Juegos

La Teoría de Juegos (TJ) es un área de la matemática aplicada que utiliza modelos para estudiar interacciones en estructuras formalizadas de incentivos (los llamados «juegos») y llevar a cabo procesos de decisión. Sus investigadores estudian las políticas óptimas así como, el comportamiento previsto y observado de agentes en juegos.

Un juego es una representación formal de una situación en la cual un número de agentes interactúan en un contexto de *interdependencia estratégica*. Esto significa que cada agente tiene un control parcial de la situación ya que, su bienestar depende no sólo de sus acciones sino, también de las acciones realizada por los otros agentes. Además, cada agente evalúa que tan bueno es realizar las distintas acciones dependiendo de lo que espera que hagan los otros agentes.

La TJ puede ser utilizada en al menos dos maneras en el diseño de sistemas multiagente:

1. Diseño del agente: la TJ puede analizar las decisiones del agente y computar la utilidad esperada para cada decisión (asumiendo que los otros agentes están actuando óptimamente de acuerdo a la TJ).

2. Diseño del mecanismo: es posible definir las reglas del ambiente (el juego que los agentes deben jugar) tal que el beneficio colectivo de todos los agentes es maximizado cuando cada agente adopta la solución de TJ que maximiza su propia utilidad.

2.6.1. Diseño de mecanismos

El Diseño de mecanismos es un área de TJ que tiene como objetivo diseñar las reglas (pagos o restricciones) de un juego (sistema multiagente) para llegar a un resultado específico. Se realiza estableciendo una estructura en la que cada jugador (agente) tiene un incentivo si se comporta como el diseñador pretende.

Formalmente un mecanismo consta de:

1. Un lenguaje para describir un política conjunta π_c: dado un conjunto de n agentes, una política conjunta π_c es una unión de políticas para cada agente $\pi_c = \pi_1, ..., \pi_n$ por lo tanto, si se tiene un conjunto de n agentes y cada agente tiene l políticas posibles que puede adoptar, entonces, existe un conjunto (posiblemente infinito) de combinaciones entre políticas individuales π_i.

2. Una regla de comportamiento (definida por el diseñador del juego): determina los balances para los agentes, dado su conjunto de políticas conjuntas.

3. Un resultado: que es una política conjunta π_c que mapea un valor numérico para cada agente, este valor es la suma de utilidades totales dada la política del agente π en la política conjunta. $\vec{f}(\pi_c) = (f(\pi_1), ..., f(\pi_n)) \in \mathbb{R}$

A continuación enumeramos algunos de los criterios que permiten definir y evaluar una regla de comportamiento en un mecanismo [30].

- *Bienestar Social*: es la suma de los pagos o utilidades de todos los agentes en una solución dada. El bienestar social mide o determina el grado de bienestar global de los agentes.

- *Pareto óptimo* : Una solución π_c es pareto eficiente o pareto óptimo si no hay otra solución π'_c en la que al menos un agente esté mejor en π'_c que en π_c, y ningún agente esté peor en π'_c que en π_c. Las soluciones que maximizan el bienestar social son un subconjunto de políticas pareto eficientes.

- *Racionalidad Individual*: La participación de un agente en un mecanismo de coordinación es racional individual, si el pago de dicho agente en la solución negociada no es menor que el pago que éste recibiría si no participará en la negociación.

- *Estabilidad*: Ningún agente debería tener incentivos para desviarse de las estrategias previamente acordadas. Por este motivo, los mecanismos deberían motivar a los agentes para que éstos se comportaran de la manera deseada.

En este trabajo, se busca conducir un conjunto agentes estratégicos a una política conjunta, pareto óptimo, de manera voluntaria con la finalidad de maximizar el bienestar social del sistema al cual pertenecen dichos agentes.

Capítulo 3

Trabajo relacionado

Este capítulo tiene por objetivo, presentar una revisión del problema de coordinación en Sistemas Multiagente (SMA), tema principal de la presente tesis. Con este propósito, se analiza el concepto de coordinación en el ámbito de los SMA así como, las razones por las cuales dicha coordinación es fundamental. Se describen y comparan dos clasificaciones de coordinación respecto a los tipos de SMA, y los principales trabajos de investigación desarrollados para dar solución a este problema.

Finalmente, se definen el enfoque sobre el cual se basa esta investigación y sus principales contribuciones, que dan solución a las limitantes que presentan los trabajos expuestos en este capítulo.

3.1. Coordinación

Retomando la definición de SMA, presentada en capítulos anteriores, se puede destacar que un SMA está compuesto por un número determinado de agentes, donde cada uno posee sus propias habilidades y objetivos, cierta especialización en cuanto a la tarea que debe realizar y un nivel limitado de conocimiento en relación directa con dicha tarea.

Si bien todo SMA posee una misión (objetivo global), su grado de cumplimiento dependerá del grado de coordinación por parte de los agentes que lo conforman respecto a los objetivos particulares que cada uno de ellos persigue. Debido a esto los

SMA propician la toma de decisiones distribuidas, lo cual permite que cada agente mantenga su autonomía para decidir que hacer y cumplir sus propósitos dentro del entorno compartido. Por lo tanto, para que un SMA cumpla su objetivo global, es necesario una exitosa coordinación entre todos los agentes del sistema.

La mayoría de las definiciones de coordinación encontradas resultan ser una conceptualización al área de Teoría de organizaciones donde, es vista como la gestión de dependencias entre actividades organizativas [25].

Bajo este enfoque, la coordinación en el ámbito de los SMA se refiere a la distribución de tareas a fin de alcanzar los objetivos comunes [26]. Tal distribución se consigue cuando los agentes pueden seleccionar las acciones apropiadas, las veces adecuadas, y en una secuencias correcta [29]. Jennings [17] asocia esta selección con un proceso de razonamiento llevado a cabo por los agentes. En dicho proceso se deben evaluar las acciones propias, y también anticipar las de los otros agentes, con la intención de procurar y asegurar que los actos de la comunidad, formada por todos los agentes, resulten coherentes.

La *coherencia* implica que las acciones de los agentes pueden realizarse sin generar conflictos entre unos y otros, es decir, logrando que los agentes se comporten como una unidad [18]. Lesser [27] sostiene que además de las selecciones de acciones y su posterior ordenamiento en una forma correcta, para coordinar también se requiere la comunicación de resultados entre los agentes, afirmando además que sin esto es imposible lograr que el grupo trabaje efectivamente. Así vista, la coordinación de actividades de los agentes comienza a ser necesaria cuando existen interdependencias, temas que se desarrolla en la sección que continúa.

3.1.1. Dependencias entre actividades de los agentes

Varios investigadores resaltan la vinculación de las dependencias o interrelaciones existentes entre las acciones o actividades a desarrollar por varios agentes con la coordinación de dichas tareas [28], [29], [9], [27], [15], [16]. Los trabajos encontrados se sostienen sobre la clasificación de interdependencias dada por Von Martial [6],

basándose en dos categorías principales:

1. *Interrelación Negativa*: Dadas dos acciones A_1 y A_2, a ser realizadas por sendos agentes, existe entre ellas una interrelación negativa si no pueden efectuarse simultáneamente. Esta imposibilidad puede deberse a la incompatibilidad de objetivos o metas perseguidas por los agentes involucrados, o a la competencia por el uso de recursos limitados. Este tipo de interrelación también recibe el nombre de interrelación conflictual.

2. *Interrelación Positiva*: Dadas dos acciones A_1 y A_2, a ser realizadas por sendos agentes, existe entre ellas una interrelación positiva si al realizar las favorecen la concreción de las tareas de otros agentes, tendiendo así a aumentar la eficiencia del sistema por no efectuar las actividades independientemente unas de otras. Las interrelaciones positivas también reciben el nombre de interrelaciones sinérgicas.

Detectar los tipos de interrelaciones existentes entre las actividades resulta importante dentro de la temática de la coordinación de tareas. Tal como afirma Ferber [9], coordinar implica articular las acciones individuales realizadas por cada agente a fin de lograr un sistema con operación de alto rendimiento. En otras palabras, analizar las interdependencias permite identificar ciertas acciones particularmente sensibles para el objetivo que se persigue en el sistema, ya sea porque sus ejecuciones simultáneas generan conflictos o porque favorecen a niveles de rendimiento individuales y globales.

Schumacher [15] sostiene que para una coordinación efectiva es necesario detectar y administrar dependencias de índole diversa dentro de un SMA. Este autor propone considerar las dependencias existentes entre los propios agentes, analizándolas desde dos puntos de vista: subjetivo y objetivo. Así, Schumacher [15] define como dependencias objetivas a las que son resultado de la composición adoptada por la organización de los agentes. También son conocidas como dependencias inter-agente.

Las dependencias subjetivas resultan del análisis del comportamiento de cada agente individual interactuándo con otros. Algunas veces se les denomina dependen-

cias intra-agente. En opinión del autor citado, ambos tipos de dependencias conviven en todo SMA.

3.1.2. Razones para coordinar

Algunas de las razones por las cuales es necesario coordinar las acciones de múltiples agentes son las siguientes [9], [17], [18],[16]:

- Los agentes individualmente no poseen las competencias requeridas para resolver el problema completo, o necesitan de los resultados que únicamente otros agentes pueden generar.
- Los recursos son limitados, por tanto sus utilizaciones deben optimizarse, evitando también los conflictos por su uso.
- Existe la necesidad de satisfacer restricciones globales para que la solución alcanzada sea considerada exitosa.
- Existen interdependencias positivas y/o negativas entre las acciones a realizarse.
- Para incrementar la eficiencia, cuando por ejemplo la información descubierta por un agente puede ser usada por otro, tal que juntos resuelven el problema más rápido.
- Para evitar el caos y la anarquía en la que pueden caer fácilmente las sociedades cuyos individuos sólo poseen metas, vistas y conocimientos locales o propios.

En general, la necesidad y la importancia de una coordinación de acciones efectiva aumentan en aquellas situaciones donde se observa una combinación de las razones arriba citadas [27].

3.2. Enfoques de coordinación en un SMA

Los trabajos de [7], [1] diferencian la coordinación bajo dos enfoques diferentes: coordinación en un enfoque descentralizado o coordinación en un enfoque centralizado.

A continuación se describen ambos enfoques.

3.2.1. Coordinación multiagente en un enfoque descentralizado

En la coordinación a nivel descentralizado los agentes explícitamente inician una comunicación y acuerdan un protocolo de comunicación. El intercambio de conocimiento se efectúa sobre la base de la comunicación directa entre los agentes. Este tipo de coordinación requiere saber con certeza quiénes son los interlocutores, dónde se encuentran y cuándo se dará tal interacción.

Se han desarrollado diversas investigaciones para alcanzar la coordinación en SMA, que cumpla con ideología de toma de decisiones a nivel distribuido. Cualquiera de ellos implementa y utiliza mecanismos, compuestos por una variedad de protocolos y estructuras, que posibilitan administrar las distintas formas de dependencias que naturalmente se dan cuando los agentes poseen objetivos interconectados, cuando comparten un ambiente común, o cuando existen recursos compartidos [2].

Todos los mecanismos de coordinación tienen propiedades y características distintas, resultando adecuados para diferentes tipos de tareas y ambientes, por lo que es imposible afirmar que exista un mecanismo universalmente bueno [2]. Básicamente se pueden agrupar en cuatro categorías, según sea su filosofía de base, a saber: estructura organizacional, contratación, planificación multiagente y negociación [9], [17], [18], [15], [16]. A continuación se describen las categorías citadas, así como las limitantes que presentan cada una de ellas.

3.2.2. Técnicas de Coordinación

Estructura Organizacional

Una estructura organizacional puede ser vista como un patrón de información y relaciones de control entre individuos. Aquí la coordinación se realiza explotando la estructura subyacente "a priori". La propia organización define implícitamente las responsabilidades de los agentes, sus capacidades, conectividad y flujo de control.

Esto define un marco para la actividad y la interacción, definiendo roles, rutas de comunicación y relaciones de autoridad.

Bajo esta categoría los mecanismos de coordinación se implementan a través de dos formas. En la primera, un agente principal planifica y distribuye tareas a los agentes subordinados, quienes pueden o no comunicarse entre sí, pero deben reportar los resultados al agente principal. Esta forma es conocida como maestro/esclavo y es apta para organizaciones jerárquicas. La segunda alternativa es la aplicada por Wulfhorst en [22], consiste en el uso de una pizarra, donde los agentes depositan mensajes o los recuperan. Es la opción para estructuras organizativas no jerárquicas.

La estructura maestro/esclavo resulta adecuada para agentes que responden al mismo tipo de organización. El uso de las pizarras presupone homogeneidad entre los agentes coordinados. Por otro lado, si los agentes involucrados son numerosos, pueden generarse severos cuellos de botella.

Contratación

Haugeneder en [8] presenta una estructura organizacional descentralizada, donde los agentes pueden asumir dos roles. El primero es desempeñado por un administrador que parte el problema en subproblemas y busca agentes a quienes contrata para que los resuelvan. Estos agentes constituyen el segundo rol, los contratistas. En este tipo de estructura cada agente puede actuar simultáneamente como administrador y contratista. El protocolo más utilizado en este caso es el de red de contratos.

Esta técnica resulta adecuada cuando la tarea puede descomponerse en subtareas que presentan un mínimo acoplamiento. No resuelve conflictos ya que presupone que los agentes intervinientes son benevolentes y no antagónicos.

Planificación multiagente

Marc en [5], efectúa una planificación multiagente descentralizada, donde a fin de evitar inconsistencias o acciones conflictivas los agentes construyen un plan multiagente que detalla todas las acciones e interacciones futuras requeridas para alcanzar

las metas. En la planificación distribuida cada agente cuenta con un modelo de los planes de los otros agentes que le permiten actualizar sus planes individuales a fin de remover conflictos. Para realizar y mantener estos modelos los agentes necesitan comunicarse entre sí.

Esta técnica resulta muy complicada debido a que, cada agente debe conocer completamente los planes de todos los demás agentes del sistema, para poder coordinar sus acciones.

Negociación

En [18] se establece que negociar implica un proceso de comunicación a través del cual un grupo de agentes alcanzan un acuerdo mutuamente aceptado respecto a algún tema conflictivo. Una parte significativa de los trabajos en el área de coordinación tratan el tema de negociación. Esta situación se debe a que muchas de las propuestas involucran algún tipo de negociación entre los agentes.

La negociación es útil para resolver conflictos en la distribución de tareas o recursos por otro lado, requiere que los agentes presenten una estructura descentralizada con comunicación directa entre ellos.

La coordinación bajo un enfoque distribuido, ha sido de gran éxito en SMA cerrados donde se tiene un control de los modelos de cada agente y de estructura general del sistema donde, la comunicación es clave fundamental para coordinar las acciones de los agentes y resolver las relaciones conflictivas.

En el caso de los SMA abiertos abiertos donde, no se tiene control de los agentes ni de la infraestructura del sistema [3] y además no existe comunicación directa entre ellos es necesario, idear otras técnicas de coordinación. Una forma de abordar el problema de coordinación en los SMA abiertos cuando no existe comunicación directa entre los agentes, es mediante un enfoque centralizado mismo que definimos en la siguiente sección.

3.2.3. Coordinación multiagente en un enfoque centralizado

En la coordinación centralizada existe un agente intermediario que recibe todos los planes parciales desde los agentes individuales. Un análisis de los mismos le permite identificar potenciales inconsistencias e interacciones conflictivas. El agente coordinador modifica los planes parciales combinándolos en un plan multiagente donde se eliminen las interacciones conflictivas.

Uno de los mecanismo de coordinación más común basado en este principio, es el mecanismo de subasta [16]. Un mecanismo de subasta clásico está compuesto por un grupo de agentes: donde uno de los agentes cumple el rol de subastador y los agentes restantes son los ofertantes, dicho mecanismo ha sido utilizado en muchas aplicaciones prácticas de ciencias computacionales que involucran la asignación de bienes, tareas y recursos.

El agente subastador oferta un recurso y los agentes ofertantes realizan peticiones siendo el ganador aquel agente que tiene una mayor necesidad (tiene la mejor oferta).

La razón por la cual la subasta se ha tornado tan popular es que es un escenario fácil de automatizar; esto la hace una buena primera opción a considerar como vía para que los agentes alcancen acuerdos entre sí, sin necesidad de comunicación directa entre ellos. Uno de los problemas que enfrenta este tipo de mecanismos es la identificación de información veraz, debido a la racionalidad de los agentes estos pueden enviar información falsa o mantener aquella información que le pudiera implicar una desventaja.

Por lo tanto, surge una línea de investigación que consiste en determinar bajo qué condiciones pueden emerger en el sistema propiedades globales interesantes que desemboquen en comportamientos uniformes. Para lograrlo se diseñan y analizan mecanismos de incentivos que sesgan el comportamiento racional de cada uno de los agentes, y así generan un comportamiento global deseado (si cada agente maximiza su utilidad local la sociedad entera mostrará un comportamiento deseable).

El problema ahora se centra en cómo identificar un mecanismo de incentivos ópti-

mo que garantice que los agentes aporten información verdadera, considerando que ellos serán sinceros si está en su interés proporcionar información cierta y completa.

Uno de los mecanismos de incentivos más común es el mecanismo propuesto en [23], consiste en que cada agente participante reciba del agente coordinador una cantidad igual a la utilidad global del sistema, por comportarse como él quiere, es decir a las suma de utilidades individuales del resto de los agentes. Esto permite alinear los objetivos de los agentes logrando un equilibrio perfecto entre ellos, ya que al reportar información verídica obtiene la utilidad global del sistema.

Este mecanismo tiene como desventaja que el agente coordinador necesita un presupuesto ilimitado para poder ofrecer a cada agente un incentivo tan elevado.

Parkes y Singh [4], intentan reducir el presupuesto del coordinador incluyendo una carga al incentivo extra de cada agente, esta carga se basa en que tan buena fue la contribución marginal de cada agente para el bienestar social del sistema, si su contribución es alta menor será la carga y si su contribución en baja la carga es mayor.

Esto le permite reducir considerablemente el presupuesto del coordinador pero, no asegura el balanceo de utilidades entre los agentes.

Hasta el momento, la idea de considerar un agente externo como coordinador de un sistema multiagente, entra en conflicto con la definición de autonomía de los agentes, debido a que los agentes deben ceder la toma de decisiones a otro agente externo.

En este trabajo se diseña e implemente un sistema de coordinación central que propicia la toma de decisiones distribuidas, el cual permite que cada agente mantenga su autonomía en la toma de decisiones y pueda cumplir sus objetivos individuales al mismo tiempo que participa en el cumplimiento de un objetivo global.

En este trabajo consideramos que cada agente puede decidir en no solo proporcionar información falsa o verdadera sí no que, puede decidir en hacer caso al coordinador o ignorarlo y seguir su política individual. Para lograr que el agente

autónomo coopere, se diseña un mecanismo de incentivos que le permita al coordinador no solo obtener información verídica de los agentes si no también, convencerlos de cambiar su política individual por una política conjunta que maximice el valor esperado del sistema sin que los agentes pierdan su racionalidad individual además de que, le permita distribuir la utilidad que los agentes reciben del sistema para balancear su presupuesto.

Capítulo 4

Modelo propuesto

Como se ha mencionado el objetivo de este trabajo es diseñar e implementar un sistema de coordinación central que propicie la toma de decisiones distribuidas, con lo que cada agente mantiene su autonomía para decidir que hacer y cumplir sus propios objetivos, pero que al mismo tiempo logra el beneficio global del sistema.

Imaginemos una central de taxis en una ciudad, cada taxista trabaja de forma particular realizando reportes periódicos de su ubicación y estado (ocupado o no) a la central. La central recibe las llamadas de los usuarios y con base en los reportes de los taxistas envía al taxi que se encuentre mas cerca de la ubicación del usuario que desea el servicio. La toma de decisiones efectiva del centro depende de los reportes verídicos de los taxistas, ya que pueden aprender a mentir sobre su ubicación o estado para verse beneficiados. Cada taxista puede decidir en trabajar por su cuenta o trabajar con el centro, aunque cuando trabaja de forma particular se ve limitado a la colonia en la que se encuentra ubicado (percepción limitada) mientras que el centro puede conseguirle clientes de otras colonias sin necesidad que el de vaya a buscarlos, pero también puede quitarle usuarios al asignárselos a otros taxistas. En escenarios como este, es necesario idear una forma de motivar a los agentes a no solo participar con el sistema si no a revelar información verídica.

4.1. Sistema de Coordinación propuesto

Para cumplir el objetivo propuesto y solucionar el problema de coordinación en un SMA abierto, donde no existe comunicación directa entre los agentes, se propone diseñar e implementar una sistema de coordinación central (ver figura 4.1). Dicho sistema define un coordinador, que recibe información individual desde los agentes, en cada periodo de decisión. En base a esta información recolectada, el coordinador propone una acción conjunta y le envía a cada agente su acción individual a manera de recomendación, bajo una política conjunta donde no solo se eliminen las interacciones conflictivas, si no que, se maximiza el valor esperado del sistema. Basado en un mecanismo de incentivos el coordinador es capaz de: 1) obtener información verídica de los agentes y 2) convencerlos de cambiar su política individual por una política conjunta, sin perder su racionalidad individual.

A continuación se describe formalmente el funcionamiento de cada componente que conforma el sistema de coordinación propuesto.

4.1.1. Sistema multiagente

El sistema de coordinación se centra en un sistema multiagente abierto formado por un conjunto $N = \{1, ..., n\}$, de n número de agentes, el sistema es no monótono debido a que cada agente $i \in N$ cumple con las siguientes características:

- Modela al entorno bajo un $MDP =< S_i, A_i, \Phi_i, R_i >$(ver capítulo 2), individual y local, por lo que el sistema es multidimensional.
- Es autónomo y actúa de forma estratégica guiado por un objetivo individual.
- Sigue su propia política π_i sin importar la de los demás.
- No tiene comunicación directa con los demás agentes del sistema.
- Tiene una percepción local del entorno es decir, existe información que un agente conoce otro no.

En SMA abiertos donde, no existe comunicación directa entre los agentes surgen conflictos debido a la forma de dependencias que se producen cuando los agentes

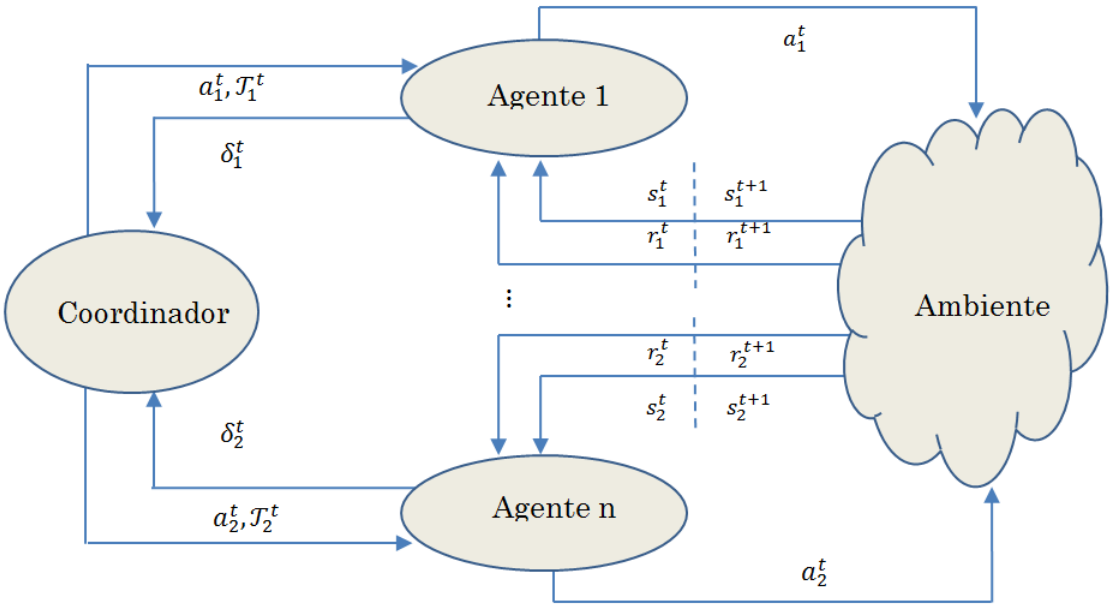

Figura 4.1: *Sistema de coordinación propuesto. Para cada paso en el tiempo t: 1) cada agente $i \in N$ reporta sus observaciones (estado privado) $\delta_i^t \in S_i$ (que puede ser falsa o verdadera) al coordinador, 2) el coordinador agrega el conjunto de observaciones $s_c^t = (\delta_1^t, ..., \delta_n^t)$ y calcula una política conjunta desde el estado s_c mediante la cual, define una acción conjunta $a_c \in \times_{i \in N} A_i$ y publica la acción individual a cada agente como recomendación 3) el coordinador realiza una transferencia T_i^t a cada agente como incentivo por reportar su información privada verdadera y por realizar la acción recomendada.*

tiene: 1) objetivos enlazados, 2) comparten un entorno en común, o 3) hay recursos compartidos. Una forma de gestionar estos conflictos es mediante un coordinador central pero esto ha contrarrestado la autonomía de los agente en la toma de decisiones. Nuestra propuesta se basa en un coordinador que propicie dicha autonomía de los agentes basado en mecanismo de incentivos que le permite convencer a los agentes a participar como unidad para lograr un fin común a partir de los objetivos individuales de cada agente.

En cada periodo de decisión t, los agentes reportan sus observaciones (estado privado) al coordinador. Debido a su autonomía en la toma de decisiones, esta información puede ser falsa o verdadera según su racionalidad es decir, los agentes pueden decidir en retener aquellas observaciones que representen alguna desventaja

para ellos. Por tal motivo, el estado s_i^t que perciben los agentes del entorno puede ser diferente al estado que reporten al coordinador δ_i^t.

Cuando los agentes reciben la recomendación del coordinación, pueden decidir en realizar dicha acción o realizar una acción de su propia política, esto permite a los agentes conservar su autonomía en la toma de decisiones.

4.1.2. Coordinador

El coordinador se diseña como un agente externo al SMA, el cual conoce los modelos individuales de cada agente y los modela bajo un MDP conjunto $MDP =< S_c, A_c, \Phi_c, R_c >$, donde:

- N es el conjunto de n agentes,
- S_c es la agregación de los estados individuales:

$$S_c = \times_{i \in N} S_i$$

- A_c es el conjunto de acciones conjuntas, dado por el producto de acciones individuales:

$$A_c = \times_{i \in N} A_i$$

  .

- Φ_c es una función de transición, dada por el producto de las funciones de transición individuales:

$$\Phi_c(s_c, a_c, s'_c) = \prod_{i \in N} \Phi_i(s'_i | s_i, a_i)$$

  donde $s_i \in S_i$, $a_i \in A_i$.

- R_c es un función de recompensa conjunta:

$$R_c = \sum_{i \in N} R_i(s_i, a_i)$$

  .

La función del coordinador es: en cada periodo de decisión t, recibe de cada agente $i \in N$ su estado privado δ_i^t, mismo que puede ser verdadero o falso. El coordinador agrega el conjunto de observaciones $s_c^t = (\delta_1^t, ..., \delta_n^t)$ y calcula una política conjunta desde el estado s_c mediante la cual, define una acción conjunta $a_c \in \times_{i \in N} A_i$ y publica la acción individual a cada agente como recomendación (ver figura 4.1).

El objetivo del coordinador es generar una política conjunta, donde no solo se eliminen las interacciones conflictivas entre los agentes, si no, que maximice el valor esperado del sistema.

$$\pi_c^*(s_c) = argmax_{a_c \in A_c} E\,[\sum_{t=t_0}^{h} R_c(s_c, a_c) | \pi_c, s_c\,] \tag{4.1}$$

Para lograr su objetivo, el coordinador necesita no solo que los agentes reporten su estado actual en cada periodo de decisión, si no que esta información sea verídica y además conseguir que los agentes sigan su recomendación. Por lo tanto, el coordinador se basa en un mecanismo que le permita incentivar a los agentes para que decidan cooperar con el sistema.

4.1.3. Mecanismo de incentivos

Dado que los agentes del SMA actúan de forma estratégica, debido a su racionalidad individual, con el objetivo de maximizar su utilidad esperada a través del tiempo. Si el coordinador les proporciona una utilidad mayor por seguir su política conjunta a la utilidad que ellos pudieran conseguir siguiendo su política individual, los agentes preferirán seguir la política del coordinador.

Por lo tanto, el coordinador se basa en el mecanismo de incentivos descrito en el Algoritmo 7, que consiste en proporcionar un incentivo a cada agente del sistema por comportarse como el coordinador quiere (ver figura 4.1). El coordinador proporciona dicho incentivo a manera de transferencia T_i a cada agente $i \in N$ en cada periodo de decisión t.

La transferencia T_i que recibe cada agente $i \in N$ del coordinador debe definirse de acuerdo a su función de recompensa ya que este incentivo puede representar-

Algoritmo 7 Mecanismo de incentivos propuesto

Para cada paso en el tiempo t:
  Cada agente $i \in N$ reporta sus observaciones $\delta_i^t \in S_i$ al coordinador.
  El coordinador agrega el conjunto de observaciones $s_c^t = (\delta_i^t, ..., \delta_n^t)$
  El coordinador define y publica una acción conjunta $a_c \in \times_{i \in N} A_i$ a los agentes como recomendación
  Para cada agente $i \in N$
    El coordinador compara los reportes del tiempo anterior s_i^{t-1}
    Si coincide **entonces**
      El coordinador realiza una transferencia $T_i(s^t)$ al agente i
    Si no coincide **entonces**
      El coordinador no realiza la transferencia al agente i

se como algo que sea significativo para el agente ya sea energía extra, capacidad de almacenamiento, ancho de banda, entre otros. En esta tesis el incentivo es monetario.

Como se ha mencionado, el objetivo de esta investigación a diferencia de otros trabajos (ver capítulo 3) es lograr que los agentes no pierdan su autonomía en la toma de decisiones, por lo tanto, como se puede observar en el algoritmo 7, la transferencia se aplica no solo por que los agentes reporten su estado privado, si no que también por realizar las acciones que recomienda el coordinador, ya que los agentes son capaces de decidir en seguir la política conjunta del coordinador o seguir su política individual, según la que más utilidad les represente.

Al comparar el reporte actual con el reporte anterior de cada agente, el coordinador puede asumir que el agente está mintiendo o que en el periodo de tiempo anterior no realizó la acción recomendada, por lo tanto, si los reportes no son consistentes el coordinador no realiza la transferencia al agente.

El valor de la transferencia que realiza el coordinador a cada agente $i \in N$ en cada periodo de decisión t, es el punto esencial del mecanismo de incentivos, ya que, al final la suma de dichas transferencias debe ser lo suficientemente buena para convencer a los agentes a adoptar la política conjunta del coordinador. Por lo tanto, inicialmente definimos un transferencia $T_i(s^t)$ dada por la siguiente ecuación:

$$T_i(s^t) = V^*(s^t) = \sum_{j \in N \setminus i} [r_j(s_j^t, \pi_c^*(s^t)) | \pi_c^*] \tag{4.2}$$

Esto indica que cada agente $i \in N$ en cada periodo de decisión t, recibirá del coordinador una transferencia que es igual al valor esperado del sistema para ese periodo de tiempo t, en el estado s menos la utilidad de i, es decir, la suma de utilidades totales que esperan recibir todos los agentes del sistema excepto el mismo, siguiendo la política del coordinador.

El propósito de utilizar dicha transferencia es alinear los objetivos de los agentes ya que, si cada agente se comporta como el coordinador quiere recibirá al final la utilidad máxima del sistema. Por lo tanto, si cada uno coopera con el objetivo global le ira mejor a los demás y por consiguiente a él mismo, entonces preferirá seguir la política del coordinador.

Uno de los inconvenientes de la ecuación 4.2 es que la cantidad que el coordinador debe pagar a los agentes, es muy alta y esto puede no ser lo más adecuado. Por lo tanto, después del análisis experimental que se presenta en los siguientes capítulos definimos una función adecuada para $T_i(s^t)$ la cual le permite al coordinador obtener información privada de los agentes y convencerlos de cambiar su política individual por una política conjunta sin invertir ninguna cantidad monetaria si no que recibiendo ganancias de los agentes. Siendo esta una de las contribuciones más importantes que presenta este trabajo de investigación.

Capítulo 5

Experimentos y resultados

En este capítulo se describe el caso de estudio sobre el cual se evalúa el sistema de coordinación propuesto en esta investigación. Se detallan las simulaciones que fueron realizadas, sus configuraciones y parámetros, así como los resultados obtenidos. Se realizan 2 etapas de simulaciones la primera con el objetivo de evaluar al coordinador sin el mecanismo de incentivos y la segunda con el objetivo de evaluar al coordinador con el mecanismo de incentivos. La serie de simulaciones en cada etapa, nos permitió diseñar una función de incentivos adecuada para disminuir la inversión del coordinador.

5.1. Caso de estudio

Para evaluar nuestro sistema de coordinación propuesto definimos un SMA abierto, donde cada agente tiene como objetivo individual recolectar los objetos que puede observar en su entorno, el cual es de tipo rejilla (que se define más adelante).

Los agentes modelan al mundo de forma independiente, tienen información local y no tienen comunicación entre ellos. Por lo tanto la interacción en un mismo entorno genera los siguientes conflictos:

- que dos agentes quieran recolectar el mismo objeto al mismo tiempo.
- que dos agentes quieran ir al mismo espacio físico al mismo tiempo

Agente	Variables de estado	Acciones	Recompensas
Agente A Agente B	$Pos(x, y)$ $\vec{Obj}(x, y)$	Arriba (↑) Abajo (↓) Izquierda (←) Derecha (→) No intervenir (o)	-1 en cada toma de decisión, donde no recolecte objeto. -5 por colisionar con pared o con otros agentes. La recompensa por recolectar un objeto varía de acuerdo al objeto (ver Figura 5.1)

Tabla 5.1: *Definición de los MDPs de los agentes del SMA. Los dos agentes tienen MDPs iguales, pero independientes, con: 2 variables de estado que indican las coordenadas de su posición en el ambiente y la de los objetos que pueden observar y 5 acciones arriba, abajo, izquierda, derecha y no intervenir o no moverse.*

Dichos conflictos pueden entorpecer la tarea de recolección de manera local para cada uno y de manera global ya que juntos pueden tardar más tiempo en recolectar todos los objetos del entorno o incluso nunca terminar. En escenarios como este la coordinación entre los agentes es fundamental, no solo en la optimización del sistema sino para lograr un bienestar social entre los agentes.

5.2. Escenario experimental

Se diseña e implementa un sistema multiagente abierto con 2 agentes, cada agente modela al entorno bajo un MDP con la información que se muestra en la tabla 5.1. El escenario experimental consiste en un ambiente abstracto conocido como "mundo de rejilla" o cuadrícula, en el cual se encuentra un conjunto de objetos que los agentes tienen como objetivo recolectar.

Se definieron dos configuraciones del escenario, Escenario 1 y Escenario 2 respectivamente, que se pueden observar en la figura 5.1. En cada una de estas configuraciones podemos encontrar: dos celdas de inicio etiquetadas con las letras A y B que corresponde a los agentes existentes (agente A y agente B) y varias celdas etiquetadas con un signo de pesos y un valor numérico que indican un objeto a recolectar y el valor de recompensa por recolectar dicho objeto.

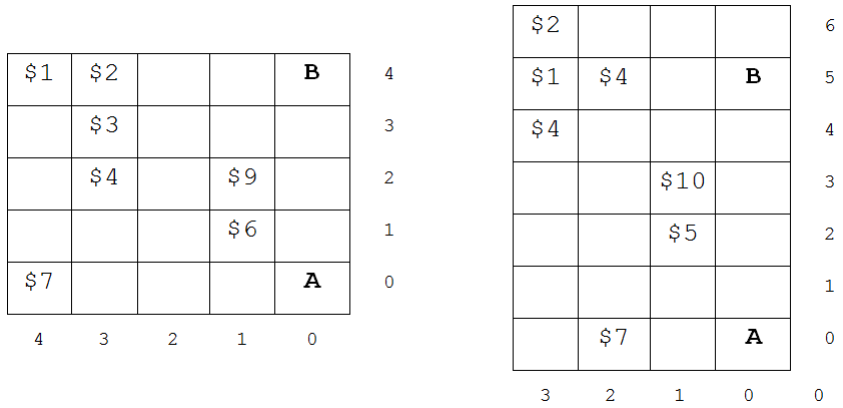

Figura 5.1: *Escenarios experimentales, mundo de rejilla, entorno del SMA.*

Cada agente cuentan con dos variables de estado en común (las cuales conforman su estado s_i) que identifican como su información privada: $Pos(x, y)$, que corresponde a las coordenadas de la celda en la que están posicionados en cada periodo de decisión, así como $\vec{Obj}(x, y)$, que corresponde a un vector de coordenadas de los objetos que observan.

La función de transición de cada agente es determinista $\Phi(s, a, s') = 1$ (ver capítulo 2), es decir la probabilidad que tiene un agente i de pasar a un estado $s_i^{'}$, al realizar una acción a_i desde un estado s_i, es siempre igual a 1.

Cada agente tiene información parcial sobre el ambiente que corresponde a la información percibida en un radio de 3 celdas, es decir pueden observar una cuadricula de 5×5 respeto a su posición, como se muestra en la figura 5.2. En esta figura, el agente A tiene una percepción marcada por el contorno rojo y el agente B una percepción marcada por el contorno azul. Esta área de percepción cambia conforme los agentes se mueven en el entorno.

El objetivo de cada agente es recolectar todos los objetos que puede percibir en su entorno por lo tanto, cuando ambos agentes perciben el mismo objeto, dicho objeto se objetivo de ambos agentes y si intentan recolectarlo al mismo tiempo (en el

mismo periodo de tiempo t) el objeto objetivo cada agente tiene el 50 % de probabilidad de quedarse con el objeto y el otro 50 % de perderlo pero, ambos reciben una penalización por colisión (definida en la tabla 5.1), lo mismo sucede en el caso de las colisiones entre los agentes.

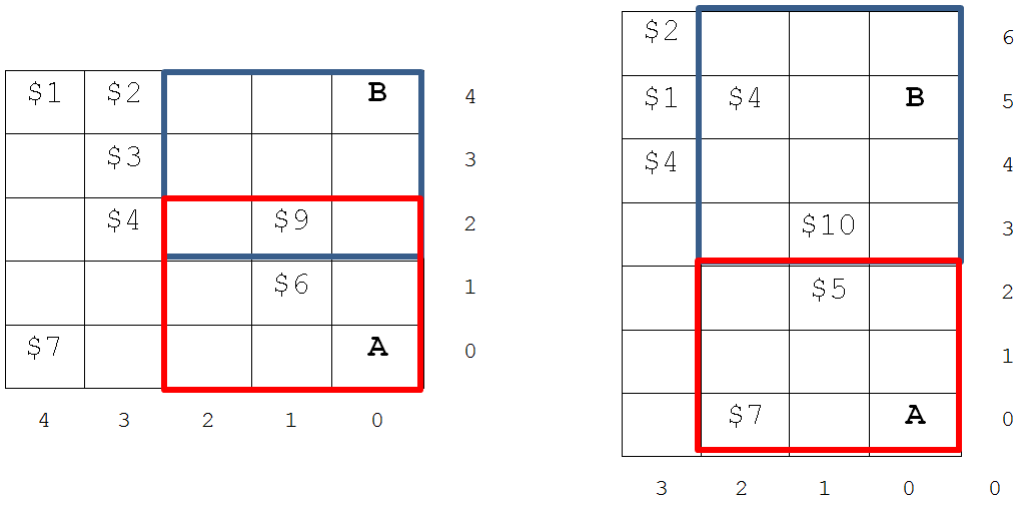

Figura 5.2: *Percepción de los agentes en el entorno, cada agente puede observar una cuadricula de* 5×5*, el agente A observa el área marcada por el contorno rojo y el agente B el área marcada por el contorno azul.*

Se definieron los dos escenarios presentados en la figura 5.1 ya que después de un análisis experimental de diversos escenarios, se identificaron como los escenarios donde los agentes eran susceptibles a mentir en cuanto a sus observaciones. Este análisis nos permitió definir la política conjunta óptima de ambos escenarios para efectos de evaluación. Dicha política se muestra en la figura 5.3.

Para efectos de evaluación, se calculo previamente bajo programación dinámica el valor máximo teórico, que puede obtener el sistema respecto a los agentes A y B siguiendo la política conjunta óptima presentada en la figura 5.3 es: para el escenario 1 de 25 (agente A=22 y agente B=3) y para el escenario 2 de 28 (agente A=19 y agente B=9).

$1←	$2←			B	4
	$3↑			↓	3
	$4↑	←	$9←	↓	2
		↓	$6←	↑	1
$7←	←	←		A	0
4	3	2	1	0	

$2↓	←			6
$1↓	$4↑	←	B	5
$4↓				4
		$10↑		3
	→	$5↑		2
	↑			1
	$7←	←	A	0
3	2	1	0	0

Figura 5.3: *Política conjunta óptima para el agente A y el agente B, para ambos escenarios.*

5.3. Descripción de los experimentos

Se diseñaron 2 etapas de experimentos con los escenarios descritos en la figura 5.1.

En la primera etapa se diseñaron 3 experimentos (experimentos 1, 2 y 3) con el objetivo de analizar el comportamiento de los agentes en el SMA modelado con y sin el coordinador propuesto, con el fin de obtener las políticas individuales, así como evaluar la autonomía de los agentes en la toma de decisiones, en dos enfoques diferentes: 1) cuando los agentes pueden decidir solo en proporcionar información falsa o verdadera y 2) cuando los agentes pueden decidir en proporcionar información falsa o verdadera y además en seguir la política del coordinador o seguir su propia política.

De la misma forma en la segunda etapa se diseñaron 3 experimentos (experimento 4, 5 y 6) con el objetivo de evaluar el mecanismo de incentivos propuesto (ver capítulo 4) y mejorar la función de transferencia de dicho mecanismo para lograr que el coordinador convenza a los agentes sin realizar una inversión monetaria.

En cada experimento consta de 50 ejecuciones de 20 mil episodios cada una, donde un episodio es una secuencia de pasos (toma de decisiones) para llegar de una estado inicia a un estado final es decir, un episodio es desde el momento en que los

agentes son posicionados en su estado inicial hasta el momento en que es recolectada el último objeto en ele entorno.

5.4. Experimento 1: Comportamiento de los agentes sin coordinador

Se tiene un SMA abierto con 2 agentes independientes, en un entorno tipo rejilla como se mencionó en la sección 5.2. Los agentes no tienen comunicación entre ellos y ambos siguen su política individual motivados por su función de recompensa que se basa en obtener mayor utilidad por recolectar la mayor cantidad de objetos en el entorno.

El objetivo de este experimento es analizar el comportamiento de los agentes en el ambiente, sin ningún tipo de coordinación, permitiendo la identificación de conflictos entre ellos y analizar como aprenden a solucionar los conflictos de forma individual.

Cada agente resuelve su MDP bajo aprendizaje por refuerzo, utilizando el algoritmo Q-Learning descrito en el capítulo 2, configurado con los siguientes parámetros: $\alpha = 0{,}1$, $\gamma = 0{,}95$ y $\epsilon = 0{,}1$, que son los valores comúnmente usados para este tipo de problemas.

Para ambos escenarios se realizaron 50 ejecuciones de 20 mil episodios cada una, de las cuales se reporta el promedio de la utilidad máxima obtenida con la política aprendida de cada agente, así como la utilidad global del sistema.

5.4.1. Resultados

Las figuras 5.4 y 5.5 muestran los resultados obtenidos (promedio de las 50 ejecuciones) al realizar este experimento en el escenario 1. La figura 5.4 muestra la utilidad obtenida por cada agente cuando el valor de penalización por colisión es de -5, mientras la figura 5.5 cuando el valor de penalización por colisión es de -10.

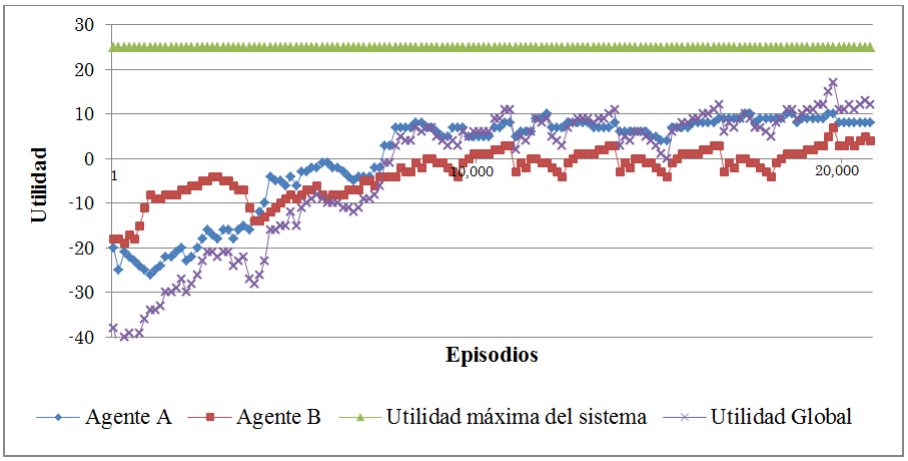

Figura 5.4: *Gráfica de aprendizaje de los agentes A y B sin coordinador respecto al escenario 1, con valor de penalización de -5.*

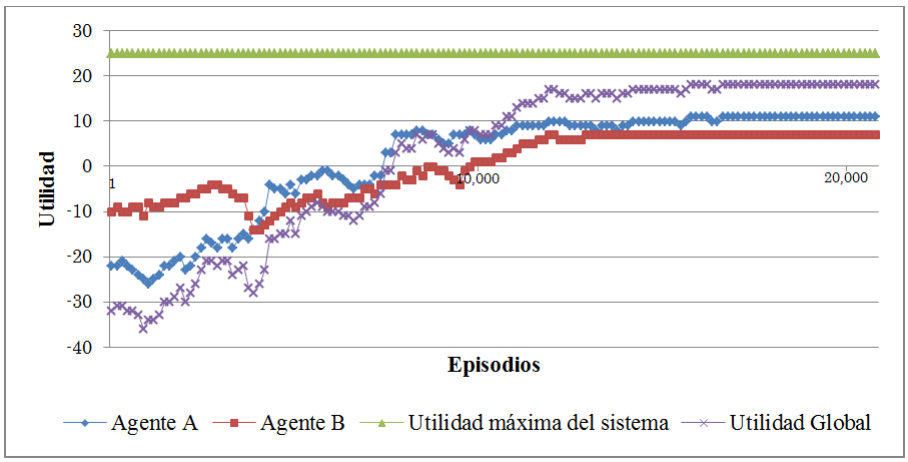

Figura 5.5: *Gráfica de aprendizaje de los agentes A y B sin coordinador respecto al escenario 1, con valor de penalización de -10.*

Analizando los resultados de este escenario, podemos notar que cuando la penalización por colisión es mayor entonces los agentes tienden a converger a una política individual y lo hacen antes de los 20 mil episodios, debido a que prefieren renunciar a ciertos objetos que son objetivos del otro agente ya que arriesgarse a ir por ellos les representa mayor pérdida que no hacerlo.

A diferencia de la gráfica donde la penalización por colisión es de -5 los agentes

Agentes	Objetos recolectados	Utilidad máxima	No. de pasos
Agente A	\$6, \$9, \$7	11	11
Agente B	\$2, \$1, \$3, \$4	7	9
Sistema	\$6, \$9, \$7, \$2, \$1, \$3, \$4	18	11

Tabla 5.2: *Valores de las políticas aprendidas por los agentes A y B, cuando el valor de penalización por colisión es de -10, en el escenario 1. Muestra los objetos recolectados por cada agente y su utilidad máxima obtenida, así como el No.de pasos que representa el número de toma de decisiones máximo.*

no logran converger en una política individual antes de los 20 mil episodios, ya que prefieren arriesgarse a colisionar con el otro agente porque tienen un 50 % de probabilidad de quedarse con el objeto que desean.

En la tabla 5.2 se muestran los valores resultantes de las políticas aprendidas para agente A y el agente B, cuando el valor de penalización por colisión es de -10, para el escenario 1.

De acuerdo con la tabla 5.2, cuando la penalización por colisión es de -10, el agente A logra converger en una política que le permite obtener una utilidad máxima de 11 en 11 pasos (toma de decisiones) y el agente B una utilidad máxima de 7 en 9 pasos, lo que permite al sistema obtener una utilidad máxima de 18 con ambos agentes, cuando la utilidad máxima esperada es de 25 como se mencionó en la sección 5.2.

De la misma forma que en el escenairo 1, las figuras 5.6 y 5.7 muestran los resultados obtenidos al realizar este experimento en el escenario 2. La figura 5.6 muestra la utilidad obtenida por cada agente cuando el valor de penalización por colisión es de -5, mientras que la figura 5.7 muestra la utilidad obtenida cuando el valor de penalización por colisión es de -10.

Analizando las resultados, correspondientes al escenario 2, se nota que cuando la penalización por colisión es mayor entonces los agentes tienden a converger en una política individual antes de los 20 mil episodios, debido a que prefieren renunciar a

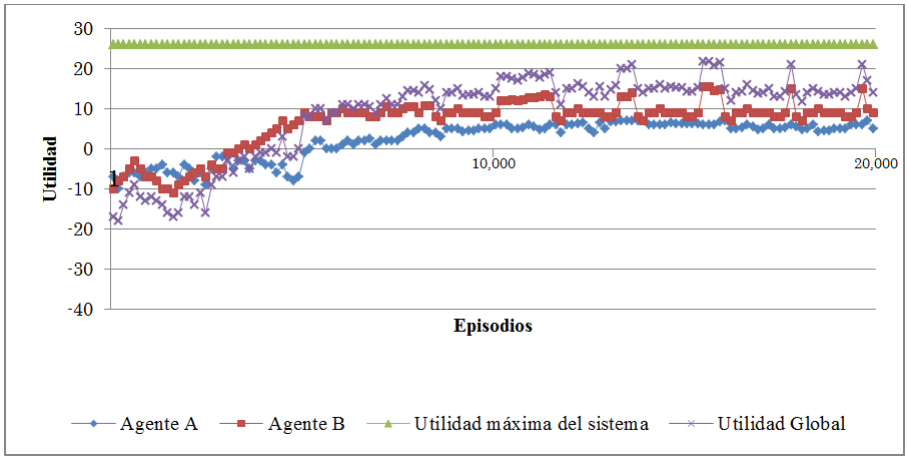

Figura 5.6: *Gráfica de aprendizaje de los agentes A y B sin coordinador respecto al escenario 2, con valor de penalización por colisión de -5.*

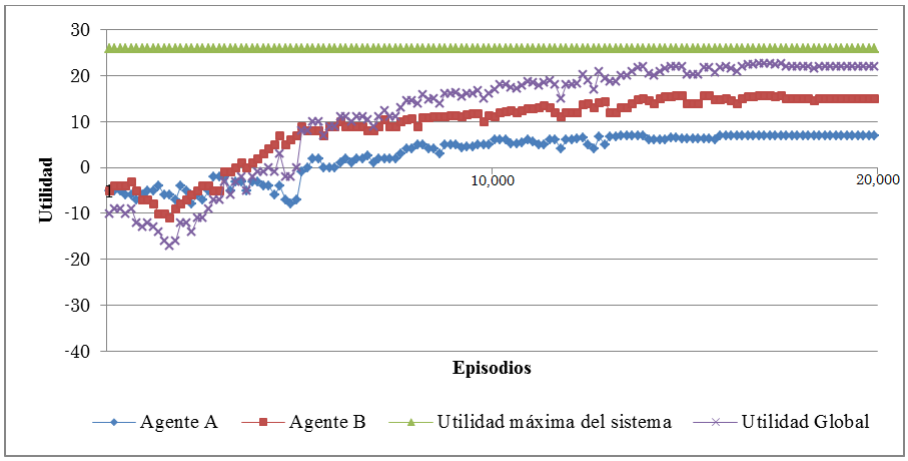

Figura 5.7: *Gráfica de aprendizaje de los agentes A y B sin coordinador respecto al escenario 2, con valor de penalización de -10.*

ciertos objetos que son objetivos del otro agente ya que arriesgarse a ir por ellos les representa mayor pérdida que no hacerlo.

A diferencia de la gráfica donde la penalización por colisión es de -5, y en donde los agentes no logran converger a una política individual antes de los 20 mil episodios, ya que prefieren arriesgarse a colisionar con el otro agente porque tienen un 50 % de probabilidad de quedarse con el objeto que desean.

Agentes	Objetos recolectados	Utilidad máxima	No. de pasos
Agente A	\$5, \$7	8	6
Agente B	\$10, \$4, \$1, \$2 \$4	16	10
Sistema	\$5, \$7, \$10, \$4, \$1, \$2, \$4	24	10

Tabla 5.3: *Valores de las políticas aprendidas por los agentes A y B, cuando el valor de penalización por colisión es de -10, en el escenario 2. Muestra los objetos recolectados por cada agente y su utilidad máxima obtenida, así como el No.de pasos que representa el número de toma de decisiones máximo.*

En la tabla 5.3 se muestra los valores resultantes de las políticas aprendidas para el agente A y el agente B, cuando el valor de penalización por colisión es de -10 para el escenario 2.

De acuerdo con la tabla 5.3, cuando la penalización por colisión es de -10, el agente A logra converger en una política que le permite obtener una utilidad máxima de 8 en 6 pasos (toma de decisiones) y el agente B una utilidad máxima de 16 en 10 pasos, lo que permite al sistema obtener una utilidad máxima de 24 con ambos agentes, cuando la utilidad máxima esperada es de 28 como se mencionó en la sección 5.2.

5.4.2. Discusión de resultados

De acuerdo a los resultados obtenidos en este experimento es importante identificar que cuando las penalizaciones por colisión (conflicto) son altas, los agentes renuncian a ciertos objetos y convergen en una política individual que para ese entorno les permite obtener la mayor utilidad. Como se observa en las tablas 5.2 y 5.3 para esta simulación, la utilidad del sistema es menor a la considerada por la utilidad máxima óptima que se definió en la sección 5.2 cuando los agentes siguen su política individual.

Retomando la esencia de esta investigación, el objetivo es optimizar el sistema, es decir lograr que los agentes se coordinen de manera óptima, sin perder su racionalidad individual y obtener el máximo valor esperado del sistema en el menor tiempo posible,

que obviamente los agentes por si solos, como lo muestran los resultados de este experimento, no pueden lograr.

5.5. Experimento 2: Agentes deciden reportar su estado falso o verdadero.

En este experimento se tiene como objetivo analizar el comportamiento de los agentes A y B cuando se implementa el coordinador propuesto, con la restricción de que los agentes obedezcan completamente las recomendaciones del coordinador, pero tengan la autonomía de decidir en reportar su estado privado falso o verdadero según les convenga, sin recibir un incentivo extra.

Como se describió en el capítulo 4, el coordinador modela al SMA bajo un $MDP_c =< S_c, A_c, \Phi_c, R_c >$ conjunto, tiene conocimiento de los MDP individuales de cada agente y en cada periodo de decisión t, el coordinador recibe el estado privado $\delta_i^t \in S_i$(que puede ser falso o verdadero) de cada agente $i \in N$. Agrega este conjunto de observaciones $s_c^t = (\delta_1^t, ..., \delta_n^t)$ y publica una acción conjunta $a_c \in \times_{i \in N} A_i$ a los agentes, siguiendo una política conjunta π_c que maximice el valor esperado del sistema.

Para este experimento, los agentes A y B, tiene dos acciones que refuerzan, reportar su estado verdadero o reportar cualquier otro estado, debido a que las acciones que el coordinador les indique realizar dependerán de lo que ellos reporten.

De la misma forma que en el experimento anterior, para ambos escenarios se realizaron 50 ejecuciones de 20 mil episodios cada una, de las cuales se reporta el promedio de la utilidad máxima obtenida con la política aprendida de cada agente, así como la utilidad global del sistema.

5.5.1. Resultados

En la figura 5.8 se muestra la gráfica de aprendizaje del agente A y del agente B para el escenario 1 y la figura 5.9 para el escenario 2, que son resultado del promedio

de las 50 ejecuciones realizadas.

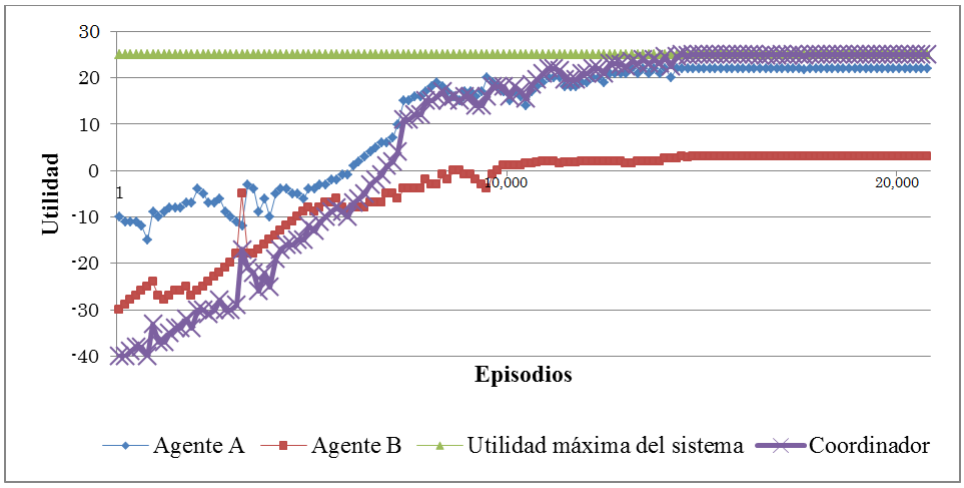

Figura 5.8: *Gráfica de aprendizaje de los agentes A y B con coordinador, respecto al escenario 1, con la restricción de que los agentes obedecen completamente al coordinador pero puede decidir en reportar su estado privado falso o verdadero, según les convenga.*

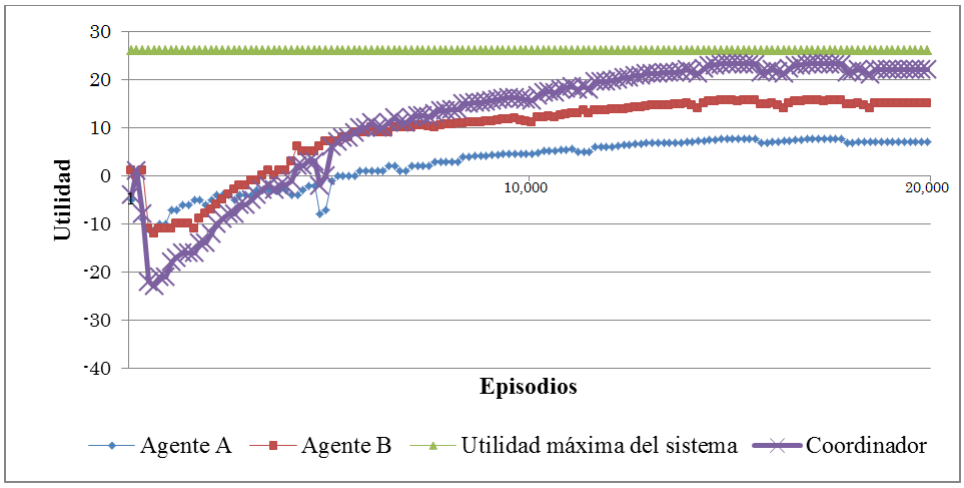

Figura 5.9: *Gráfica de aprendizaje de los agentes A y B con coordinador, respecto al escenario 2, con la restricción de que los agentes obedecen completamente al coordinador pero puede decidir en reportar su estado privado falso o verdadero, según les convenga.*

Escenario	Agentes	Objetos recolectados	Utilidad máxima	No. de pasos
1	Agente A	\$6, \$9, \$4, \$7	22	8
	Agente B	\$2, \$1, \$3,	3	6
	Sistema	\$6, \$9, \$7, \$2, \$1, \$3, \$4	25	8
2	Agente A	\$5, \$7	8	6
	Agente B	\$10, \$4, \$1, \$2 \$4	16	10
	Sistema	\$5, \$7, \$10, \$4, \$1, \$2, \$4	24	10

Tabla 5.4: *Valores de las políticas aprendidas por los agentes A y B, cuando obedecen al coordinador pero pueden decidir en reportar su estado privado falso o verdadero, para ambos escenarios*

Como podemos observar, en la gráfica de la figura 5.8 en el escenario 1 los agentes convergen rápidamente al valor máximo esperado del sistema, a diferencia del escenario 2 de la gráfica de la figura 5.9. La tabla 5.4 muestra los valores resultantes a las políticas en las que logran converger los agentes A y B, para ambos escenarios.

De acuerdo a los resultados obtenidos y recordando las características de los escenarios, podemos notar que para el escenario 1, el agente A prefiere reportar su estado verdadero debido a que el coordinador le puede dar a ganar una mayor utilidad que si lo hubiera hecho solo o hubiera reportado un estado falso. Así mismo el agente B prefiere reportar siempre su estado verdadero debido a que si no reporta que ve a los objetos \$1, \$2, \$3 y \$4 el coordinador le hubiera indicado quedarse en su posición de inicio sin ganar nada, ya que esto era lo más conveniente para el sistema. Por lo tanto prefiere decir la verdad y tener una utilidad de 3 como lo indica la tabla 5.4.

Para el escenario 1 el coordinador converge en la política óptima debido a que los agentes convergen a reportar su estado verdadero en cada periodo de decisión además, no solo se obtiene la utilidad máxima del sistema si no que, el número de pasos se reduce considerablemente, es decir los agentes requieren menos tiempo en recolectar todos los objetos.

Mientras que en el escenario 2, pasa completamente lo contrario ambos agentes

prefieren mentir al coordinador. Si el agente B miente, es decir no reporta que observa a los objeto \$2, \$4, \$1, \$4, el coordinador lo va a enviar por el objeto \$10 y posiblemente por el objeto \$5, después de recolectarlos puede reportar a los demás objetos y el coordinador lo mandará por ellos ya que estará más cerca que el agente A, y de esta forma el agente B puede tener mayor utilidad, reportando un estado falso. De la misma manera el agente A prefiere no reportar al objeto \$7 antes de recolectar al objeto \$5, de esta forma podrá quedarse con ambos objetos. Por lo tanto, el coordinador converge en una política que no es la política óptima definida en en la sección 5.2, ya que los agentes no reportan un estado verdadero en cada periodo de decisión.

5.5.2. Discusión de resultados

En el escenario 1 los agentes convergen en la utilidad máxima óptima además de reducir el tiempo que necesitan para cumplir el objetivo global del sistema (recolectar todos los objetos del entorno) debido a que reportan su estado verdadero en cada periodo de decisión. A diferencia del escenario 2 que es un claro ejemplo de que debido a la naturaleza estratégica de los agentes, éstos pueden estar incentivados a reportar información falsa para poder obtener una mayor utilidad de forma independiente.

Entonces, debido a que en este experimento los agentes están obligados a obedecer al coordinador reportar un estado falso es la única forma que les permite cumplir su objetivo individual, que es obtener mayor utilidad al recolectar la mayor cantidad de objetos.

Con estos resultados surge la pregunta que da pie al siguiente experimento. Si los agentes no están obligados a obedecer al coordinador, ¿seguirán reportando información falsa?

5.6. Experimento 3: Agentes deciden reportar estado privado falso o verdadero y en obedecer o no al coordinador

El objetivo de este experimento es analizar el comportamiento de los agentes A y B con el coordinador propuesto cuando ambos agentes tienen la autonomía en decidir no solo en reportar su estado privado (falso o verdadero), si no que también tienen autonomía para seguir la política conjunta que el coordinador recomienda o seguir su propia política.

De la misma forma que el experimento anterior, el coordinador modela al SMA bajo un $MDP_c =< S_c, A_c, \Phi_c, R_c >$ conjunto, tiene conocimiento de los MDP individuales de cada agente y en cada periodo de decisión t, el coordinador recibe el estado privado $\delta_i^t \in S_i$ (que puede ser falso o verdadero) de cada agente $i \in N$, agrega este conjunto de observaciones $s_c^t = (\delta_1^t, ..., \delta_n^t)$ y publica una acción conjunta $a_c \in \times_{i \in N} A_i$ a los agentes, siguiendo una política conjunta π_c que maximice el valor esperado del sistema. Los agentes reciben dicha recomendación y pueden decidir en ejecutar la acción recomendada o realizar la acción que les indique su política individual.

Para este experimento los agentes A y B modelan al entorno bajo el MDP descrito en la tabla 5.1, con un valor de penalización de -10, con el objeto de identificar si siguen la política definida en el experimento 1 o siguen la política del coordinador.

De la misma forma que en el experimento anterior, para ambos escenarios se realizaron 50 ejecuciones de 20 mil episodios cada una, de las cuales se reporta el promedio de la utilidad máxima obtenida con la política aprendida de cada agente, así como la utilidad global del sistema.

5.6.1. Resultados

En la figura 5.10 se muestra la gráfica de aprendizaje del agente A y el agente B, para el escenario 1 y la figura 5.11 para el escenario 2 (los escenarios se describen en la figura 5.1), que son resultado del promedio de las 50 ejecuciones realizadas.

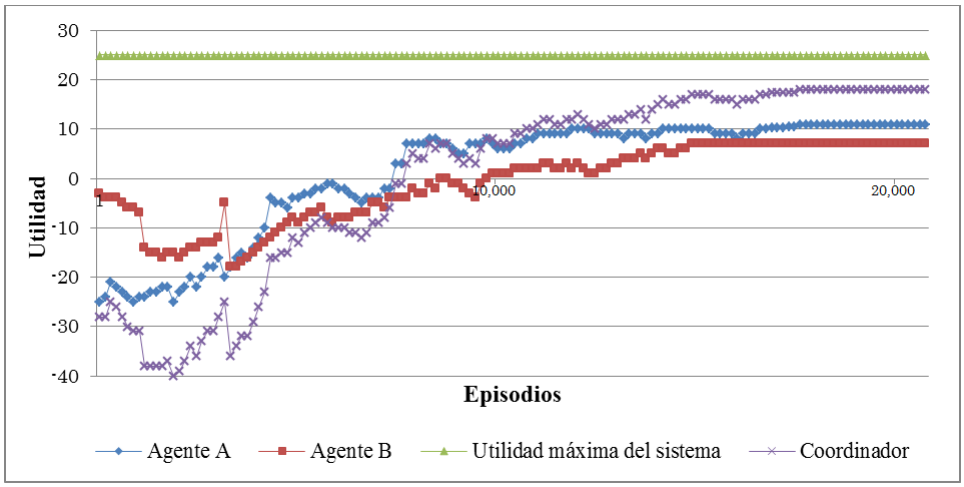

Figura 5.10: *Gráfica de aprendizaje de los agentes A y B con coordinador, respecto al escenario 1, cuando los agentes deciden en reportan su estado privado falso o verdadero al coordinador y en seguir la política conjunta del coordinador o seguir su política individual.*

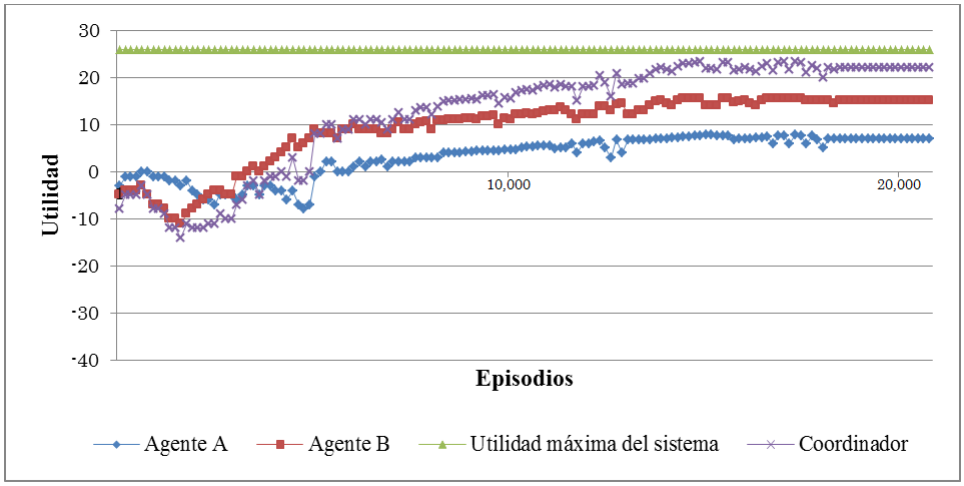

Figura 5.11: *Gráfica de aprendizaje de los agentes A y B con coordinador, respecto al escenario 2, cuando los agentes deciden en reportan su estado privado falso o verdadero al coordinador y en seguir la política conjunta del coordinador o seguir su política individual.*

Las gráficas de las figuras 5.10 y 5.11 muestran que los agentes no convergen en

Escenario	Agentes	Objetos recolectados	Utilidad máxima	No. de pasos
1	Agente A	\$6, \$9, \$7	11	11
	Agente B	\$2, \$1, \$3, \$4	7	9
	Sistema	\$6, \$9, \$7, \$2, \$1, \$3, \$4	18	11
2	Agente A	\$5, \$7	8	6
	Agente B	\$10, \$4, \$1, \$2 \$4	16	10
	Sistema	\$5, \$7, \$10, \$4, \$1, \$2, \$4	24	10

Tabla 5.5: *Valores de las políticas resultantes en la que convergen los agentes A y B, cuando los agentes deciden en reportan su estado privado falso o verdadero al coordinador y en seguir la política conjunta del coordinador o seguir su política individual.*

la utilidad máximo esperada del sistema (linea verde), como el caso del escenario 1 del experimento anterior (sección 5.5). La tabla 5.5 muestra los valores resultantes a las políticas en las que convergen los agentes A y B en este experimento, para ambos escenarios.

Podemos observar que los resultados obtenidos en la tabla 5.5, son iguales a los de las tablas 5.2 y 5.3, obtenidos en el experimento 1 (sección 5.4), donde no existe coordinador, lo que nos indica que los agentes A y B, tanto en el escenario 1 como en el escenario 2, prefieren seguir su propia política que obedecer al coordinador independientemente de mentir o decir la verdad respeto a su estado privado.

En el escenario 1 el agente B prefiere seguir su política individual ya que con ella puede obtener una utilidad de 7 a diferencia de la que el coordinador le puede hacer ganar que es de 3, por otro lado el agente A podría ganar 22 si obedece al coordinador pero al seguir su propia política el agente B, entonces al agente A debe seguir su política individual para por lo menos asegurar una utilidad de 11.

El mismo caso se presenta en el escenario 2 ya que el agente B prefiere seguir su política individual, con la que puede obtener una utilidad de 16 a diferencia de la que el coordinador le puede hacer ganar que es de 9. Por otro lado el agente A

podría ganar 19 si obedece al coordinador pero al seguir su propia política el agente B, entonces al agente A debe seguir su política individual para por lo menos asegurar una utilidad de 8.

5.6.2. Discusión de resultados

De acuerdo con los resultados obtenidos en este experimento, si los agentes tienen la libertad de poder decidir los agentes en obedecer al coordinador o seguir su política individual, van a decidir en seguir su política individual sin importar si el reporte de su estado privado es falso o verdadero, ya que no necesitan manipular al coordinador si pueden ignorarlo.

Al seguir un agente su política individual obliga a los demás agentes a hacer lo mismo, ya que de lo contrario pueden salir perjudicados.

Finalmente, al poder los agentes decidir en seguir la política del coordinador o seguir su propia política sin que haya incentivos, es como si no hubiera coordinador ya que los agentes siguen su política individual sin importar lo que el coordinador recomiende.

Debido a estos resultados, surge la interrogante que hemos planteado en esta investigación, ¿cómo hacer que el coordinador motive a los agentes a cambiar su política individual por una política conjunta, sin que pierdan su autonomía en la toma de decisiones? como fue el caso del experimento 2 (sección 5.5). El siguiente experimento da solución a esta interrogante.

5.7. Experimento 4: Evaluación del mecanismo de incentivos

En este experimento se plantea que el coordinador utilice el mecanismo de incentivos que se propuso en esta investigación, descrito en el algoritmo 7 (ver capítulo 4) con una transferencia definida por la ecuación 4.2. Es decir el coordinador realiza una transferencia $T_i(s^t)$ en cada paso en el tiempo t a cada agente $i \in N$, por repor-

tar su estado verdadero y por realizar la acción que le recomienda, que es igual a la recompensa inmediata esperada, para ese tiempo t de todos los agentes excepto el mismo, siguiendo la política conjunta del coordinador.

De la misma forma que el experimento anterior, el coordinador modela al SMA bajo un $MDP_c =< S_c, A_c, \Phi_c, R_c >$ conjunto, tiene conocimiento de los modelos individuales de cada agente y en cada periodo de decisión t, el coordinador ejecuta el mecanismo de incentivos definido en el algoritmo 7.

Para este experimento los agentes A y B, modelan al entorno bajo el MDP descrito en la tabla 5.1, con un valor de penalización por colisión de -5.

Para ambos escenarios, se realizaron 50 ejecuciones de 20 mil episodios cada una, de las cuales se reporta el promedio de la utilidad máxima obtenida con la política aprendida de cada agente, así como la utilidad global del sistema.

5.7.1. Resultados

En la figura 5.12 se muestra la gráfica de aprendizaje del agente A y el agente B, para el escenario 1 y la figura 5.13 para el escenario 2 (los escenarios se describen en la figura 5.1), resultado del promedio de las 50 ejecuciones realizadas.

Las figuras 5.12 y 5.13 muestran que los agentes convergen en el máximo valor teórico esperado del sistema, es decir gracias al mecanismo de incentivos los agentes prefieren reportar al coordinador su estado privado verdadero y seguir las acciones recomendadas por él, ya que al hacerlo reciben un incentivo extra que les representa una mayor utilidad.

La tabla 5.6 muestra los valores resultantes a las políticas en las que convergen los agentes A y B, en este experimento, para ambos escenarios.

Analizando los resultados obtenidos en la tabla 5.6, se observa que en el escenario 1 el agente B renuncia a recolectar un objeto y por lo tanto, a obtener una recompen-

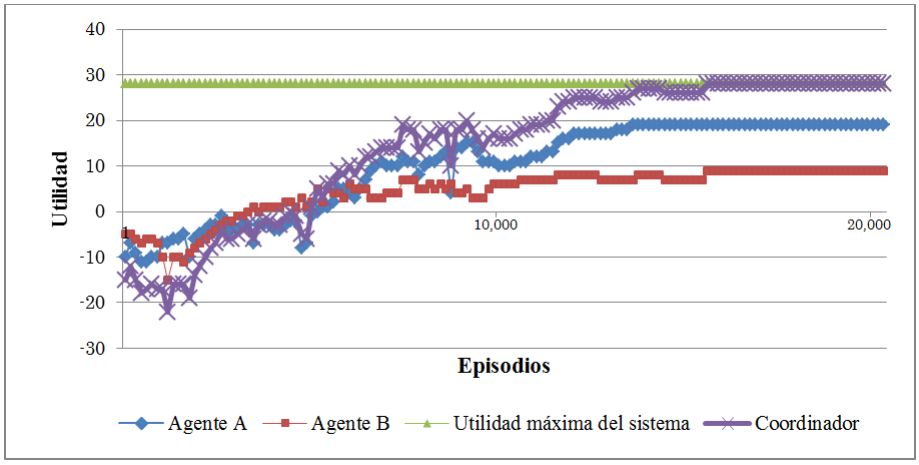

Figura 5.12: *Gráfica de aprendizaje de los agentes A y B con coordinador, respecto al escenario 1, cuando el coordinador utiliza el mecanismo de incentivos propuesto con una transferencia definida por la ecuación 4.2..*

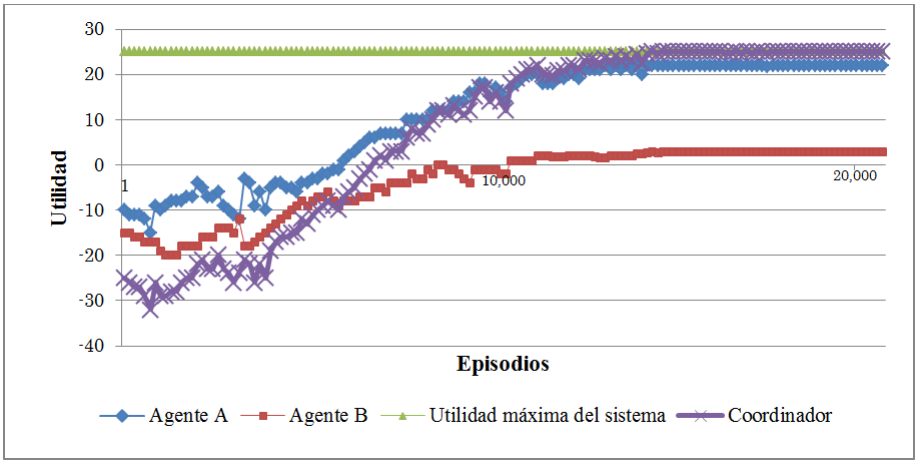

Figura 5.13: *Gráfica de aprendizaje de los agentes A y B con coordinador, respecto al escenario 2, cuando el coordinador utiliza el mecanismo de incentivos propuesto con una transferencia definida por la ecuación 4.2.*

sa de 3 siguiendo la política conjunta del coordinador que una de 7 con su política individual, ya que al reportar su estado privado verdadero y seguir la recomendación del coordinador recibe un incentivo extra igual al utilidad máxima obtenida por el agente A que es de 22 recibiendo una utilidad de 25 que es mucho mejor que el 7 que podía conseguir el solo.

Escenario	Agentes	Objetos recolectados	Utilidad máxima	No. de pasos	Suma de transferencias
1	Agente A	\$6, \$9, \$4, \$7	22	8	3
	Agente B	\$2, \$1, \$3,	3	6	22
	Sistema	\$6, \$9, \$7, \$2, \$1, \$3, \$4	25	8	-25
2	Agente A	\$7, \$5, \$10	19	6	9
	Agente B	\$4, \$1, \$2 \$4	9	6	19
	Sistema	\$5, \$7, \$10, \$4, \$4 \$1, \$2, \$4	28	6	-28

Tabla 5.6: *Valores de las políticas en la que convergen los agentes A y B, cuando el coordinador implementa el mecanismo de incentivos propuesto con una transferencia definida por la ecuación 4.2*

De la misma forma, el agente A prefiere seguir la política del coordinador debido a que con ella puede ganar una utilidad de 22 más 3 que es el incentivo extra que recibe del coordinador, que es igual a lo que recibe el agente B, en total recibe una utilidad de 25 que es mucho mayor al 11 que recibía siguiendo su política individual.

Así mismo, para el escenario 2 el agente B renuncia a una utilidad de 16 por una utilidad de 9 que puede obtener al seguir la política del coordinador, debido a que al hacerlo recibe un incentivo extra de 19.

Los resultados de este experimento nos permiten definir que los agentes siguen la política del coordinador ya que de lo contrario no hubieran alcanzado una política óptima en menos tiempo (menor número de pasos), como se vió en experimento 1 del capítulo anterior, cuando el valor de penalización de colisión fue de -5 como es el caso de este experimento.

5.7.2. Discusión de resultados

Los resultados de este experimento, muestran que el mecanismo de incentivos le permite al coordinador lograr que los agentes cambien su política individual por la política conjunta que él recomienda, logrando la utilidad máxima del sistema en un

menor tiempo, sin que los agentes pierdan su racionalidad individual pero pagando un costo muy elevado. Ya que al realizar la transferencia base definida por la ecuación 4.2, el coordinador paga a los agentes un valor igual al la utilidad máxima esperada del sistema.

Aunque con esta transferencia logramos el objetivo general de esta investigación es necesario, analizar los resultados obtenidos y tratar de balancear la utilidad del sistema para evitar que el coordinador invierta demasiado en la coordinación de los agentes.

Este balance se centra en identificar que aunque el agente B en ambos escenarios recibe una utilidad menor al seguir la política del coordinador que su política individual, el agente A (para ambos escenarios) recibe una utilidad mucho mayor al seguir la política del coordinador que su política individual. Por lo tanto, podríamos tratar de balancear la utilidad del agente A, para evitar que el coordinador realice un pago elevado a los agentes.

Este análisis da paso al siguiente experimento, que nos permite balancear la utilidad del sistema y reducir el presupuesto del coordinador de forma satisfactoria.

5.8. Experimento 5: Ajuste del presupuesto del coordinador

Para este experimento es necesario analizar el presupuesto del coordinador obtenido en el experimento anterior al realizar la transferencia propuesta inicialmente descrita por la ecuación 4.2 (ver capítulo 4), resumido en la tabla 5.6.

De acuerdo con la ecuación 4.2, el coordinador realiza un incentivo total igual a la utilidad máxima esperada del sistema, ya que al seguir ambos agentes la política conjunta del coordinador, reciben una transferencia total que es igual a la suma de utilidades que reciben los demás agentes excepto el mismo por cada periodo de decisión.

Por lo tanto, este experimento consiste en dividir la transferencia base definida en la ecuación 4.2, en algún valor x, con el fin de disminuir el gasto que realiza el coordinador sin cambiar los resultados logrados en el caso de experimentos anterior (ver sección 5.7), mismo que cumple con el objetivo base de dicha investigación. Por lo tanto, la transferencia del mecanismo de incentivos definido en el Algoritmo 7, se modifica por la siguiente ecuación:

$$T_i(s^t) = x \sum_{j \in N \setminus i} V_j(s^t, a^t) \tag{5.1}$$

Se evalúan ambos escenarios modificando el valor de la transferencia $T_i(s^t)$ del mecanismo de incentivos, por la ecuación 5.1, con los siguientes valores para $x = \{\frac{1}{2}, \frac{1}{3}, \frac{1}{4}\}$, es decir la transferencia base de la ecuación 4.2 se divide a la mitad, en un tercio y en un cuarto.

5.8.1. Resultados

Los resultados obtenidos se muestran en la tabla 5.7, para el escenario 1 y en la tabla 5.8 para el escenario 2.

La tabla 5.7 muestra la utilidad máxima obtenida (UM) y la suma de transferencias totales (ST) que obtiene cada agente cuando recibe una trasferencia $T_i(s^t)$ en cada periodo de decisión t, definida por la ecuación 5.1, para cada valor de $x = \{\frac{1}{2}, \frac{1}{3}, \frac{1}{4}\}$, para el escenario 1. Mientras que la tabla 5.8 lo hace para el escenario 2.

Se puede observar que cuando el incentivo que realiza el coordinador a cada agente, en cada periodo de decisión, es igual a un transferencia de 0 ($x = 0$) el agente A converge en una utilidad máxima de 11 y el agente B a una de 7, mismas que corresponden a sus políticas individuales, es decir, si no reciben ningún incentivo extra del coordinador no siguen su política conjunta, como se definió en experimentos pasados.

Cuando $x = \frac{1}{4}$, los agentes convergen en la política del coordinador; el agente A converge en una utilidad máxima de 22 más el incentivo extra proporcionado por el

	$x = 0$		$x = \frac{1}{4}$		$x = \frac{1}{3}$		$x = \frac{1}{4}$		$x = 1$	
Escenario 1	***UM***	***ST***	***UM***	***ST***	***UM***	***ST***	***UM***	***ST***	***UM***	***ST***
Agente A	11	0	22	0.75	22	1	22	1.5	22	3
Agente B	7	0	3	5.5	3	7.33	3	11	3	22
Gasto del coordinador		0		6.25		8.33		12.5		25

Tabla 5.7: *Tabla de resultados. Muestra la utilidad máxima obtenida (UM) y suma de transferencias totales (ST) que obtiene cada agente cuando recibe una trasferencia $T_i(s^t)$ en cada periodo de decisión t, definida por la ecuación 5.1, para cada valor de $x = \{\frac{1}{2}, \frac{1}{3}, \frac{1}{4}\}$. Para el escenario 1.*

	$x = 0$		$x = \frac{1}{4}$		$x = \frac{1}{3}$		$x = \frac{1}{4}$		$x = 1$	
Escenario 2	***UM***	***ST***	***UM***	***ST***	***UM***	***ST***	***UM***	***ST***	***UM***	***ST***
Agente A	8	0	8	4	8	5.33	19	4.5	19	9
Agente B	16	0	16	2	16	2.66	9	9.5	9	19
Gasto del coordinador		0		6		8		14		28

Tabla 5.8: *Tabla de resultados. Muestra la utilidad máxima obtenida (UM) y suma de transferencias totales (ST) que obtiene cada agente cuando recibe una trasferencia $T_i(s^t)$ en cada periodo de decisión t, definida por la ecuación 5.1, para cada valor de $x = \{\frac{1}{2}, \frac{1}{3}, \frac{1}{4}\}$. Para el escenario 2.*

coordinador de 0.75 y el agente B converge en una utilidad máxima de 3 más un incentivo extra proporcionado por el coordinador de 5.5 que le suma una utilidad total de 8.5 que es mayor al 7 que pudo haber obtenido siguiendo su propia política.

Por lo tanto, podemos observar que para el escenario 1, basta con proporcionar a los agentes un cuarto de la transferencia base definida en la ecuación 5.1, para que los agentes prefieran cambiar su política individual por la política conjunta del coordinador. Reduciendo el presupuesto del coordinador de 25 (con la ecuación 4.2) a 6.25 (con la ecuación 5.1).

A diferencia del escenario 1, en el escenario 2, como se muestra en la tabla 5.8, los agentes A y B cambian su política individual por la política conjunta del coordinador, hasta que $x = \frac{1}{2}$, ya que de lo contrario el agente B recibe una utilidad menor que la que podría recibir siguiendo su política individual. Por lo tanto, en este escenario el gasto del coordinador se reduce solo a la mitad de 28 a 14.

5.8.2. Discusión de resultados

Estos experimentos nos muestran que es posible reducir el presupuesto del coordinador y seguir logrando resultados deseados en los agentes, si encontramos un valor adecuado para x en cada escenario.

Una forma de hacerlo es evaluar con diversos valores para x, como se realizó en estos experimentos, identificando el que más reduzca el presupuesto del coordinador sin afectar los resultados deseados en los agentes. Sin embargo esta estrategia de prueba y error puede ser demasiado tardada.

Analizando los resultados de las tablas 5.7 y 5.8, podemos afirmar que el agente B en ambos escenarios cambiará su política individual por la política conjunta del coordinador, siempre que este le proporcione un incentivo extra que sea igual a la utilidad máxima que puede obtener siguiendo su política individual menos la utilidad máxima que puede obtener siguiendo la política del coordinador más un ϵ.

Por ejemplo en el escenario 1, si el agente B recibe un incentivo de $7-3+\epsilon = 4+\epsilon$, el agente B tendrá una recompensa total de $3+4+\epsilon = 7+\epsilon$, que por mínimo que sea el valor de ϵ, es mayor que 7, misma que es la utilidad que podría conseguir siguiendo su política individual.

En el siguiente experimento formalizamos esta afirmación para definir un valor para x que le permita al coordinador cumplir su objetivo con la mínima inversión.

5.9. Experimento 6: Redistribución de utilidades

Esta sección final de experimentos es la más importante y la que define el éxito de esta investigación. Formalizando el análisis realizado en la discusión de resultados del experimento anterior (sección 5.8.2), tenemos la siguiente afirmación:

Un agente i cambiará su política individual π_i por la política conjunta del coordinador π_c, si y solo si, la suma de transferencias T_i que recibe del coordinador en cada periodo de decisión t, es igual a la utilidad máxima obtenida siguiendo su política individual π_i menos la utilidad máxima obtenida siguiendo la política conjunta del coordinador π_c más una valor mínimo ϵ.

$$\sum_{t=0}^{h} T_i(s^t) = \sum_{t=0}^{h} [r_i(s^t, \pi^*(s^t))|\pi^*] - \sum_{t=0}^{h} [r_i(s^t, \pi_c^*(s^t))|\pi_c^*] + \epsilon \tag{5.2}$$

Donde h representa un horizonte finito de toma tiempo (ver capitulo 2).

Considerando que

$$T_i(s^t) = x \sum_{j \in N \setminus i} V_j(s^t, a^t)$$

$$\sum_{t=0}^{h} x_i \sum_{j \in N \setminus i} V_j(s^t, a^t) = \sum_{t=0}^{h} [r_i(s^t, \pi^*(s^t))|\pi^*] - \sum_{t=0}^{h} [r_i(s^t, \pi_c^*(s^t))|\pi_c^*] + \epsilon$$

$$x_i \sum_{t=0}^{h} \sum_{j \in N \setminus i} V_j(s^t, a^t) = \sum_{t=0}^{h} [r_i(s^t, \pi^*(s^t)) | \pi^*] - \sum_{t=0}^{h} [r_i(s^t, \pi_c^*(s^t)) | \pi_c^*] + \epsilon$$

Entonces

$$x_i = \frac{\sum_{t=0}^{h} [r_i(s^t, \pi^*(s^t)) | \pi^*] - \sum_{t=0}^{h} [r_i(s^t, \pi_c^*(s^t)) | \pi_c^*] + \epsilon}{\sum_{t=0}^{h} \sum_{j \in N \setminus i} V_j(s^t, a^t)} \tag{5.3}$$

Por lo tanto, en este experimento, el coordinador utiliza el mecanismo de incentivos propuesto (ver capítulo 4), con una transferencia $T_i(s^t)$ a cada agente i, calculada mediante la ecuación 5.1 y 5.3, con 3 valores de $\epsilon = \{0.1, 0.01, 0.001\}$.

De la misma forma, que en los experimentos pasados, en este experimento los agentes A y B, modelan al entorno bajo el MDP descrito en la tabla 5.1, con un valor de penalización por colisión de -5.

Se realizaron 50 ejecuciones de 20 mil episodios cada una, de las cuales se reporta el promedio de la utilidad máxima obtenida con la política aprendida de cada agente, así como la utilidad global del sistema. Para ambos escenarios.

5.9.1. Resultados

En la tabla 5.9 se muestran los resultados obtenidos para el escenario 1, y la tabla 5.10 los resultados obtenidos para el escenario 2.

En las tablas 5.9 y 5.10 para ambos escenarios, el coordinador tiene una gasto total negativo, lo que indica que siguiendo el mecanismo de incentivos propuesto, con una transferencia basada en las ecuaciones 5.1 y 5.3, el coordinador no realiza un gasto, si no que recibe una utilidad.

Analizando el escenario 1 cuando $\epsilon = 0.1$, el agente A obtiene una utilidad máxi-

Escenario 1	**Utilidad máxima**	**No. de pasos**	**Suma total de transferencias**		
			$\epsilon = 0.1$	$\epsilon = 0.01$	$\epsilon = 0.001$
Agente A	22	8	-9.65	-9.74	-9.749
Agente B	3	6	5.35	5.26	5.251
Gasto del coordinador			-4.3	-4.48	-4.498

Tabla 5.9: *Resultados del experimento 6 para el escenario 1, muestra la utilidad máxima que obtiene cada agente del entorno, su No. de pasos (número de toma de decisiones) total, el incentivo total recibido por el coordinador (suma de transferencias) para cada valor de ϵ.*

Escenario 2	**Utilidad máxima**	**No. de pasos**	**Suma total de transferencias**		
			$\epsilon = 0.1$	$\epsilon = 0.01$	$\epsilon = 0.001$
Agente A	19	6	-8.14	-8.21	-8.221
Agente B	9	6	8.1	8.01	8.001
Gasto del coordinador			-0.04	-0.2	-0.22

Tabla 5.10: *Resultados del experimento 6 para el escenario 2, muestra la utilidad máxima que obtiene cada agente del entorno, su No. de pasos (número de toma de decisiones) total, el incentivo total recibido por el coordinador (suma de transferencias) para cada valor de ϵ.*

ma del entorno de 22, pero recibe un incentivo total del coordinador de -9.65, quedando con una utilidad total de 12.35, que es mayor al 11 que obtenía siguiendo su política individual, entonces el agente A recibe una utilidad extra de 1.35 más de lo que hubiera conseguido solo. El agente B obtiene una utilidad máxima del entorno de 3, pero recibe un incentivo total del coordinador de 5.35, quedando con una utilidad total de 8.35, que es mayor al 7 que obtenía siguiendo su política individual. Entonces el coordinador al quitarle al agente A, 9.65 y al agente B darle 5.35 se queda con una utilidad de 4.3, por lo que no sólo coordina al sistema si no que gana por hacerlo.

Entre menor sea el valor de ϵ mayor será la ganancia del coordinador. El agente A, en ambos escenarios, es el que obtiene una mayor utilidad siguiendo la política conjunta del coordinador que su política individual sin ningún incentivo extra del coordinador.

Por lo tanto, la suma de transferencias que el coordinador le da es negativa lo que le obliga a pagar al coordinador por hacerle ganar más, con esta ganancia el coordinador compensa al agente B que es el más perjudicado al seguir la política conjunta del coordinador, motivándolo a ceder, logrando así un bienestar social entre los agentes ya que ambos están contentos con su utilidad final.

5.9.2. Discusión de resultados

Los resultados que muestran las tablas 5.9 y 5.10, indican que ambos agentes siguen la política conjunta del coordinador ya que convergen en la utilizada máxima esperada del sistema. Dicha política conjunta es generada por el coordinador, cuando todos los agentes reportan su estado privado verdadero en cada periodo de decisión.

Ambos agentes renuncian a su política individual aprendida, en el experimento 1 (ver sección 5.4), debido a que reciben del coordinador un incentivo total (suma de transferencias) que aun siendo pequeño, en el caso del agente A, les permite obtener una utilidad total mayor que la que obtienen con su política individual.

El mecanismo de incentivos propuesto, y la transferencia calculada en esta sección

nos permite distribuir las utilidades de los agentes, es decir el que gana más paga una cuota por hacerlo y el que gana menos recibe un incentivo como compensación, de esta forma se alinean los objetivos de los agentes logrando un bienestar social (definido en el capítulo 2).

Capítulo 6

Conclusiones y trabajo futuro

En este capítulo, se detallan los resultados relevantes que se obtuvieron a lo largo de esta investigación, así como las afirmaciones que surgen con base en los datos experimentales. También se exponen las posibles líneas de investigación a futuro que surgen a partir de los resultados obtenidos en este trabajo.

6.1. Conclusiones

La necesidad actual de resolver problemas complejos que involucran diversas tareas, ha posicionado a los sistemas multiagente como uno de los temas de investigación más relevantes en los últimos años.

Un SMA esta compuesto por un número determinado de agentes donde, cada uno posee sus propias habilidades y objetivos, cierta especialización en cuanto a la tarea que debe realizar y un nivel limitado de conocimiento en relación directa con dicha tarea.

Si bien todo SMA posee una misión (objetivo global), su grado de cumplimiento dependerá del grado de colaboración por parte de los agentes que lo conforman respecto a los objetivos que cada uno de ellos persigue. Entonces, el éxito de un SMA en la solución de un problema, depende esencialmente de la coordinación entre los agentes.

Una coordinación efectiva es esencial si diferentes agentes autónomos quieren alcanzar sus objetivos dentro de un SMA. Dicha coordinación requiere gestionar las diferentes formas de dependencia que se producen cuando los agentes tienen: 1) objetivos enlazados, 2) comparten un entorno común, ó 3) hay recursos compartidos.

El objetivo de la coordinación, es lograr una integración y un ajuste armonioso de los esfuerzos de trabajo individual, en beneficio de una meta común. La comunicación entre los agentes, es clave para la coordinación coherente de la toma de decisiones distribuidas en un SMA. En escenarios donde no existe comunicación directa entre los agentes, caso común de los SMA abiertos, es necesario idear una forma de coordinación, donde la comunicación directa entre los agentes no sea relevante.

Una forma de coordinar SMA abiertos, donde no existe comunicación directa entre los agentes es mediante un coordinador central, el cual, a partir de información recibida desde los agentes pueda definir lineas de acción coherentes para cada agente. Sin embargo, definir un coordinador central contrarresta la autonomía de los agentes en la toma de decisiones distribuidas.

En este trabajo, se diseño un sistema de coordinación central que propicia la toma de decisiones distribuidas, lo cual permite que cada agente mantenga su autonomía en la toma de decisiones y cumpla sus propósitos individuales al mismo tiempo que cumple con un objetivo global, como parte de un sistema.

Para conservar su autonomía bajo una coordinación centralizada, cada agente debe poder decidir en proporcionar toda su información (información verdadera) o retener aquella que le represente alguna desventaja (información falsa) y decidir en considerar la política conjunta del coordinador central o seguir su propia política.

Cuando un agente tiene la autonomía solo en decidir proporcionar información verdadera o retener aquella que le represente alguna desventaja (información falsa), decidirá proporcionar información falsa siempre que esto represente una ventaja mayor que proporcionar información verdadera, de esta forma podrá seguir cumpliendo su objetivo individual sin importar el de los demás.

Cuando un agente tiene la autonomía en decidir proporcionar información verdadera o retener aquella que le represente alguna desventaja (información falsa) y además en seguir la política del coordinador o seguir su propia política, decidirá seguir su propia política siempre que esta represente una mayor utilidad para él.

La política conjunta del coordinador depende de su nivel de percepción, entre mayor numero de agente reportando su percepción local mayor sera la percepción global del coordinador. Entonces, el coordinador es capaz de definir una política conjunta óptima que maximice el valor esperado del sistema siempre y cuando los agentes proporcionen información verídica porque reportes falsos minimizan su percepción global .

Debido a la racionalidad de los agentes, su comportamiento siempre está sesgado por seguir la política que le proporcione una mayor utilidad. En este trabajo de investigación se definió un novedoso mecanismo de incentivos que le permite al coordinador no sólo lograr una política conjunta que maximice el valor esperado del sistema, en un menor tiempo (menor número de pasos en el tiempo) si no también, alinear los objetivos de los agentes. Logrando la optimización del sistema y el bienestar social entre los agentes.

El mecanismo de incentivos propuesto y la transferencia definida tras la experimentación le permite al coordinador distribuir las utilidades de los agentes es decir, el que gana más paga una cuota por hacerlo y el que gana menos recibe un incentivo como compensación. De esta forma el coordinador no solo coordina al sistema, si no que también recibe una utilidad por hacerlo.

6.2. Trabajo futuro

Las líneas principales de trabajo futuro que nos deja esta investigación, son:

- Evaluar el SMA propuesto con un mayor número de agentes.
- Llevar el modelo del coordinador a un nivel de aprendizaje, logrando que no se

necesite tener información *a priori* de los modelos individuales de cada agente (MDPs), si no que a través de la interacción y de la información que le reporten, pueda aprenderlos. De esta forma, el coordinador podría trabajar con SMA desconocidos para él.

Bibliografía

[1] Omicini A. y Zambonelli F. Coordination for internet aplication development. *Autonomous Agents and Multi-Agent Systems*, 2, Kluwer:251–269, 1999.

[2] Excelente-Toledo C. y Jennings N. The dynamic selection of coordination mechanisms. *Autonomous Agents and Multi-Agent Systems*, 9, Kluwer:55–85, 2004.

[3] Hewitt C. Offices are open systems. *ACM Transactions on Information and System Security*, 4:271–286, 1986.

[4] Parkes C. D. y Satinder S. Optimal coordinated planning amongst self-interested agents with private state. *In Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UIA)*, págs. 55–62, 2006.

[5] Marc F., Fallah-Seghrouchni A., y Degirmenciyan I. Distributed coordination based on temporal planning for tactical aircraft simulation. *Autonomous Agents and Multi-Agent Systems*, págs. 1060–1061, 2003.

[6] Von Martial F. *Coordinating Plans of Autonomous Agents*. Springer-Verlag, 1992.

[7] Cabri G., Leonardi L., y Zambonelli F. The impact of the coordination model in the design of mobile agent applications. *In Proc. Annual International Conference Computer Software and Applications*, Institute of Electrical and Electronics Engineers, CS Press, 1998.

[8] Haugeneder H. y Steiner D. Cooperating agents concepts and applications. *In Jennings and Wooldridge Edit. Agent Tecnology: Fundations, Applications, and Markets, Springer*, págs. 175–202, 1998.

[9] Ferber J. *Multi-Agent Systems. An Introduction to Distributed Artificial Intelligence.* Addison-Wesley, 1999.

[10] Watkins C. J. Learning from delayed rewards. phd thesis, king's college, cambridge, uk. 1989.

[11] Watkins C. J. y Dayan P. Q-learning. *Machine Learning*, 8:279–292, 1992.

[12] Sycara K. Multiagent systems. *AI Magazine*, 19(2):79–92, 1998.

[13] Huhns M. y Stephens L. Multiagent systems and societies of agents. *G. Weiss ed. Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence, The MIT Press*, págs. 79–120, 1999.

[14] Klein M. Enabling robust open multi-agent systems project proposal. 4, 2000.

[15] Schumacher M. Objective coordination in mas engineering. *Lecture Notes in Artificial Intelligence, Springer-Verlag*, 2039:33–49, 2001.

[16] Wooldridge M. *An Introduction to MultiAgent Systems.* Wiley Publishing, 2nd edition, 2009. ISBN 0470519460, 9780470519462. 3, 63.

[17] Jennings N. *Coordination Techniques for Distributed Artificial Intelligence.* In O'Hare and Jennings Edit. Foundations of Distributed Artificial Intelligence, Wiley, 1996.

[18] Jennings N., Nwana H., y Lee L. Coordination in software agent systems. *BT Technology Journal*, l4(4):79–88, 1996.

[19] Jennings N., Sycara K., y Wooldridge M. A roadmap to cigent research and development. *Autonomous Agents and Multi-Agent Systems*, l:275–306, 1998.

[20] Bellman R. Dynamic programming and stochastic control processes.information and control. 1(3):228–239, 1958.

[21] Bourne R., Excelente-Toledo C., y Jennings N. Run-time selection of coordination mechanisms in multi-agent systems. *In Proc. of the 14th European Conference on Artificial Intelligence ECAI-2000*, págs. 348–362, 2000.

[22] Wulfhorst R., Nakayama L., y Vicari R. A multiagent approach for musical interactive systems. *In Proc. ACM Autonomous Agents and Multiagent Systems AAMAS*, págs. 584–591, 2003.

[23] Groves T. Incentives in teams. *Ecomometrica*, 41:617–631, 1973.

[24] Jaakkola T., Jordan M., y Singh S. On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation*, 6:1185–1201, 1994.

[25] Malone T. y Crowston K. The interdisciplinary study of coordination. 1994.

[26] Dignum V. y Dignum F. Task ad social coordination in agent organizatios. *In Proc. ACM Autonomous Agents and Multiagent Systems AAMAS*, págs. 1183–1184, 2005.

[27] Lesser V. Reflections on the nature of multi-agent coordination and its implications for an agent architecture. *Autonomous Agents and Multi-Agent Systems AAMAS*, l:89–111, 1998.

[28] Chen W. y Decker K. Coordination mechanism for dependency relations among multiple agents. *In Proc. ACM Autonomous Agents and Multiagent Systems AAMAS*, 2002.

[29] Chen W. y Decker K. Managing multi-agent coordination, planning and scheduling. *In Proc. ACM Autonomous Agents and Multiagent Systems AAMAS*, 2004.

[30] Sandholm T. W. *Distributed Rational Decision Making, Multiagent Systems. A Modern Approach to Distributed Artificial Intelligence.* The MIT Press, 1999.