INSTITUTO POLITÉCNICO NACIONAL

CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN

Sistema Multiagente para el Análisis de Asociación de Datos en Fuentes Distribuidas

TESIS

Que para obtener el grado de:

MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:

Euler Hernández Contreras

Director:
Dr. Leonid Borisovitch Sheremetov

Co-Director:
Dr. Oleksiy Pogrebnyak Boleslavovich

México, D.F.

Julio de 2007

INSTITUTO POLITÉCNICO NACIONAL

Centro de Investigación en Computación

Sistema Multiagente para el Análisis de Asociación de Datos en Fuentes Distribuidas

TESIS

Que para obtener el grado de:
MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

P R E S E N T A
Euler Hernández Contreras

Director:
Dr. Leonid Borisovitch Sheremetov

Co-Director:
Dr. Oleksiy Pogrebnyak Boleslavovich

México, D. F. Julio de 2007

Agradecimientos

A mis Papás, *los Profesores Ricardo Hernández Cruz* y *Carmen Mercedes Contreras Martínez*, por estar conmigo en todo el tiempo, por su apoyo, por preocuparse por mi, por su comprensión, por motivarme a seguir adelante, sus exhortaciones y por supuesto sus oraciones en todo tiempo. Les Amo!!!

No pueden faltar mis hermanos porque han sido un ejemplo para mi, a cada uno les doy las gracias porque al menos he estado en sus pensamientos y en sus oraciones, les quiero con todo mi corazón: *Bertha, Elizabeth, Abigail, May, Neyla, Ricardo* y *Hugo*.

A mis sobrinos mayores gracias por su apoyo: *Omar, Yanira, Zuri e Ilse*; para el resto de ellos (*Isaí, Ricardo, Nael, Jonathan, Denise, Neyla, Brian, Lis, Hugo, Gerson, Hanna* incluyendo a *Janya* y *Jasiel*) espero que éste trabajo sea ejemplo para ustedes.

Gracias a todos los que han estado conmigo incluyendo a todos mis amigos, compañeros y colegas de la escuela: *Ernesto, Benjamín, Rolando, Iván* y *Lidya*.

Al *Dr. Leonid Borisovitch Sheremetov* por darme el privilegio de trabajar este proyecto, por su grande paciencia para conmigo y llevarme a la culminación del mismo; de igual manera al *Dr. Oleksiy Pogrebnyak Boleslavovich* por sus aportaciones al proyecto de tesis.

No puede pasar por alto el *Instituto Politécnico Nacional* y el *Centro de Investigación en Computación* por la oportunidad de realizar mi estudio de posgrado, gracias a todos mis maestros.

A mis *Sinodales* por su aportación tan valiosa en el presente trabajo.

A *CONACYT* por la beca otorgada en el tiempo que estuve estudiando.

Pero, sobre todas las cosas a *Ti* mi *Señor* ***JESÚS*** por estar conmigo en todo este tiempo. Sólo tú eres digno de alabanza y adoración porque has sido mi refugio, mi consuelo, mi sanador, porque todo el conocimiento y la ciencia son tuyas, gracias porque moriste en la cruz del calvario para darme Salvación. Mientras viva mi corazón y mi ser será tuyo.

Contenido

Glosario **vii**

1 Introducción **1**
1.1 Planteamiento del Problema 1
1.2 Objetivos 3
1.2.1 Objetivo General 3
1.2.2 Objetivos Específicos 3
1.3 Organización del documento 3

2 Estado de Arte **5**
2.1 Introducción a la minería de datos 5
2.2 Sistemas que generan Reglas de Asociación 8
2.3 Agentes y Plataformas de Agentes 10
2.3.1 Agentes de Software 10
2.3.2 Sistemas multiagentes 11
2.3.3 Ambientes de desarrollo 12
2.3.3.1 Agent Builder 12
2.3.3.2 Madkit 13
2.3.3.3 JATLite 13
2.3.3.4 Jade 14
2.3.3.5 Zeus 14
2.3.3.6 CAPNET 15

3 Marco Teórico **17**
3.1 Métodos básicos de clasificación y cluster análisis 17
3.1.1 Clasificación 17
3.1.2 Cluster Análisis 17
3.2 Medidas de Asociación 19
3.3 Métodos de agrupación 20
3.4 Minería de datos 21
3.4.1 Definición de Minería de Datos 21

3.4.2 Tareas de la Minería de Datos 22
3.5 Técnicas de minería de Datos 23

4 Reglas basadas en correlación 26
4.1 Introducción 26
4.2 Reglas de Asociación, definiciones 27
4.3 Ejemplos 29
4.4 Reglas de Asociación basadas en Correlación 33
4.5 Otros tipos de Reglas de Asociación 35
4.6 Metodología de generación de reglas de asociación basadas en correlación usando el paradigma de agentes 37

5 Análisis y diseño del SMA para el análisis de datos de series de tiempo y extracción de reglas de asociación 40
5.1 Análisis y diseño del SMA 40
5.1.1 Análisis 41
5.1.1.1 Casos de Uso 41
5.1.1.2 Modelo de Agentes (MA) 41
5.1.1.3 Modelo de Servicios (MS) 41
5.1.1.4 Modelo de Interacciones (MI) 41
5.1.2 Diseño 51
5.1.2.1 Modelo de Agentes (MA) 51
5.1.2.2 Modelo de Servicios (MS) 51
5.1.2.3 Modelo de Interacciones (MI) 52
5.1.2.4 Modelo SMA 52

6 Resultados Experimentales 61
6.1 Introducción 61
6.2 Descripción del caso de estudio 61
6.3 Resultados Experimentales 62

7 Conclusiones 70
7.1 Resultados 70
7.2 Contribuciones 72
7.3 Conclusiones 72
7.4 Trabajo Futuro 73

Bibliografía 74

Lista de Figuras

2.1 *Modelo de referencia de FIPA* 14
2.2 *Estructura de CAPNET* 15
2.3 *Arquitectura de CAPNET* 16

4.1 *Pasos generales de la metodología para generar reglas de asociación* . 38

5.1 *Diagrama de Casos de Uso del SMA* 47
5.2 *Diagrama de clases del SMA* 54
5.3 *Diagrama de Actividades del Servicio LeerDatos* 55
5.4 *Diagrama de Actividades del Servicio SeleccionarDatos* 55
5.5 *Diagrama de Actividades del Servicio VisualizarDatos* 56
5.6 *Diagrama de Actividades del Servicio DefinirCondiciónReglas* 56
5.7 *Diagrama de Actividades del Servicio GenerarMatrizCorrelación* . . . 57
5.8 *Diagrama de Actividades del Servicio GenerarReglasAsociación* 57
5.9 *Diagrama de Interacciones del SMA* 58
5.10 *Diagrama de Colaboración del SMA* 59
5.11 *Arquitectura de Smart-Agua* 59
5.12 *Arquitectura del Sistema Multi-Agente Propuesto* 60

6.1 *Muestra las series de tiempo contenidos en archivos Excel* 64
6.2 *Interfaz donde se muestra gráficamente las series de tiempo seleccionadas* 66
6.3 *Visualización de resultados de asociaciones entre los pozos* 69

Lista de Tablas

4.1 *Transacciones de una Base de datos* 30
4.2 *1-itemset en la Base de datos* 31
4.3 *2-itemset en la Base de datos* 31
4.4 *3-itemset en la Base de datos* 32
4.5 *4-itemset en la Base de datos* 32
4.6 *Reglas de Asociación con 1-item consecuentes de 3-itemsets* 33
4.7 *Reglas de Asociación con 2-item consecuentes de 3-itemsets* 33
4.8 *Regla de Asociación para con* $\{A, C\}$ 33
4.9 *Regla de Asociación para con* $\{B, C\}$ 34
4.10 *Regla de Asociación para con* $\{B, E\}$ 34
4.11 *Regla de Asociación para con* $\{C, E\}$ 34
4.12 *Una simple Base de Datos* 36

5.1 *Descripción Agente Datos* 42
5.2 *Descripción Agente Correlación* 42
5.3 *Descripción Agente GraficaST* 43
5.4 *Descripción Agente GeneradorRA* 43
5.5 *Descripción del servicio LeerDatos* 44
5.6 *Descripción del servicio SeleccionarDatos* 44
5.7 *Descripción del servicio VisualizarDatos* 45
5.8 *Descripción del servicio DefinirCondiciónReglas* 45
5.9 *Descripción del servicio GenerarMatrizCorrelación* 46
5.10 *Descripción del servicio GenerarReglasAsociación* 46
5.11 *Descripción de la interacción SolicitarRegistro* 48
5.12 *Descripción de la interacción SolicitarDatos* 48
5.13 *Descripción de la interacción SeleccionarDatos* 49
5.14 *Descripción de la interacción GraficarDatos* 49
5.15 *Descripción de la interacción GenerarMatrizCorrelación* 50
5.16 *Descripción de la interacción GenerarReglasAsociación* 50

6.1 *Valores Lingüísticos* 63

6.2 *Series de tiempo seleccionadas* 65
6.3 *Datos iniciales* 65

7.1 *Tabla comparativa de las metodologías* 72

Glosario

API: Del inglés Application Programming Interface. Interfaz de Programación de Aplicaciones, es un conjunto de funciones y procedimientos o métodos, que ofrece cierta librería para ser utilizado por otro software como una capa de abstracción. representa un interfaz de comunicación entre componentes software.

CAPNET: Plataforma de Desarrollo de Sistemas Multi-Agente, el cual fue desarrollado por el IMP, basado en las especificaciones FIPA y escrito completamente en el Framework de .NET de Microsoft.

CORBA: Del acrónimo del inglés Common Object Request Broker Architecture. Es un estándar que establece una plataforma de desarrollo de sistemas distribuidos facilitando la invocación de métodos remotos bajo un paradigma orientado a objetos; fue definido y está controlado por el Object Management Group (OMG) que define la API, el protocolo de comunicaciones y los mecanismos necesarios para permitir la interoperabilidad entre diferentes aplicaciones escritas en diferentes lenguajes y ejecutadas en diferentes plataformas, lo que es fundamental en computación distribuida.

DIMETER ADVISOR : Es un sistema experto desarrollado en 1980 por Schlumber Doll Research para apoyar al análisis de datos recolectados en la explotación de aceite.

DM: Es el acrónimo del inglés Data Mining. Minería de Datos , engloba un conjunto de técnicas encaminadas a la extracción de conocimiento procesable implícito en las bases de datos de las empresas. Las bases de la minería de datos se encuentran en la inteligencia artificial y en el análisis estadístico. Mediante los modelos extraídos utilizando técnicas de minería de datos se aborda la solución a problemas de predicción, clasificación y segmentación.

FIPA: Acrónimo del inglés Foundation for Intelligent Physical Agents. Fundación para los Agentes Físicos Inteligentes; es una organización de estándares en Computación de la IEEE, la cual promueve tecnologías basadas en agentes y su interoperabilidad con otras tecnologías.

ILP: Del acrónimo en inglés Inductive Logic Programming. La Programación Lógica Inductiva se puede ver como la intersección entre el aprendizaje automático y la programación lógica. En programación lógica inductiva se puede combinar resultados empíricos y métodos inductivos de aprendizaje bajo la representación de lógica de primer orden para así inducir conceptos representados por medio de un programa lógico.

IMP: Acrónimo de Instituto Mexicano del Petróleo. Es un centro de investigación en México dedicado al área petrolera, creado el 23 de agosto del 1995, cuyos objetivos principales son realizar investigación y desarrollo tecnológico, así como servicios especializados orientados a las necesidades estratégicas y operativas de Petróleos Mexicanos (Pemex); el IMP entrega soluciones integrales innovadoras y desarrolla recursos humanos especializados con un enfoque de calidad, oportunidad y precios competitivos.

KDD: Es el acrónimo del inglés Knowledge Discovery in Databases. Descubrimiento de Conocimiento en Bases de Datos, se define como el extracto no-trivial de información implícita, desconocida, y potencialmente útil de datos.

PEMEX: Petróleos Mexicanos, es la compañía estatal mexicana que se encarga de la explotación de los recursos energéticos (principalmente petróleo y gas) en territorios mexicanos. PEMEX es la única compañía que explota el petróleo en México, la cual actúa bajo la dirección de la Secretaría de Energía.

KQML: Acrónimo del inglés Knowledge Query and Manipulation Language. Lenguaje de Consulta y Manipulación de Conocimiento, es uno de los principales estándares de comunicación entre agentes. Consiste de tres capas: La capa de contenido, que tiene presente el contenido actual del mensaje. La capa de comunicación, codifica los mensajes describiendo los parámetros de bajos niveles de comunicación y la capa de mensajes que determina el tipo de interacciones que se pueden tener con un agente; su función principal es identificar los protocolos que son usados para la entrega de mensajes y suplir los actos comunicativos al contenido.

RDF: Es el acrónimo del inglés Resource Description Framework; como su nombre lo indica es un framework para describir e intercambiar metadatos. Es una DTD (definición del tipo de documento) de XML, es decir, una aplicación de metadatos que utiliza XML a fin de proporcionar un marco estándar para la interoperabilidad en la descripción de contenidos web.

RMI: Java Remote Method Invocation. Es un mecanismo ofrecido en Java para invocar un método remotamente. Al ser RMI parte estándar del entorno de ejecución Java, usarlo provee un mecanismo simple en una aplicación distribuida

que solamente necesita comunicar servidores codificados para Java; por medio de RMI un programa Java puede exportar un objeto. A partir de esa operación este objeto está disponible en la red esperando conexiones en un puerto TCP; un cliente puede entonces conectarse e invocar métodos.

SIH: Sistemas Inteligentes Híbridos, denotan a los sistemas software que emplean, en paralelo, una combinación de modelos de inteligencia artificial, métodos y técnicas de éstos subcampos tales como: Sistemas difusos expertos, redes neuronales evolutivas, algoritmos genéticos, etc.

XML: Es el acrónimo del inglés de eXtensible Markup Language (Lenguaje de marcas extensible), es un metalenguaje extensible de etiquetas desarrollado por el World Wide Web Consortium (W3C). Es una simplificación y adaptación del SGML y permite definir la gramática de lenguajes específicos (de la misma manera que HTML es a su vez un lenguaje definido por SGML). XML es una manera de definir lenguajes para diferentes necesidades.

Capítulo 1

Introducción

En este primer Capítulo se hace la introducción de la tesis, estableciendo los objetivos específicos y general de la misma, planteando la problemática a resolver y finalmente se presenta la organización de la tesis.

1.1 Planteamiento del Problema

El interés del uso de las tecnologías de información para apoyar las decisiones de las actividades operacionales, descansa en el hecho de que el personal operativo de compañías petroleras como PEMEX diariamente tienen que resolver problemas mediante el análisis de un volumen considerable de datos y del empleo de su experiencia adquirida en el campo; es por ello que en la industria del gas y el aceite ha surgido el interés de aplicar herramientas potentes, robustas e inteligentes para el análisis, procesamiento e interpretación en tiempo real de grandes cantidades de información para optimización de procesos.

La perforación de pozos es un proceso industrial complejo donde la industria del gas y el aceite ha introducido tecnologías innovadoras tales como los Sistemas Inteligentes Híbridos (SIH) que integran diferentes aspectos de inteligencia artificial debido a la capacidad de manejar con éxito las complejidades del mundo real como la imprecisión, incertidumbre y vaguedad. Un problema muy común en la producción de hidrocarburos, es el control de agua, ya que éste puede reflejar una reducción de los costos y un aumento en la producción.

El agua siempre está presente en los yacimientos y en asociación con el aceite, afectando positiva o negativamente la producción de un campo; el exceso de agua representa un problema urgente y complejo, con muchas soluciones potenciales que

dependen de la detección oportuna así como de obtener el diagnóstico más certero del problema. Por lo tanto, la solución apropiada para el control de la producción de agua sintetiza la información sobre el tipo de producción, las características del yacimiento y de los fluidos producidos, las historias de producción y perforación, las características y estado de las instalaciones, etc.

A pesar de que la industria petrolera ha sido criticada por moverse a paso de tortuga con respecto a las tecnologías informáticas, fue una de las primeras en adoptar y aplicar sistemas expertos. Como ejemplos clásicos se mencionan los sistemas DIPMETER ADVISOR y PROSPECTOR. Las nuevas generaciones de sistemas expertos fueron desarrolladas por Halliburton y Schlumberger (XERO, WaterCase) [Bailey et al,2000] [Shahab,2005].

PEMEX a través del IMP desarrollan un sistema experto denominado Smart-Agua, el cual pretende emplear experiencias en yacimientos naturalmente fracturados (presentes en escenarios mexicanos) en un sistema informático inteligente, cuya arquitectura es mostrada en el Capítulo 5. Smart-Agua, utiliza técnicas de minería de datos para el análisis de series de tiempo para la obtención de reglas de asociación, módulo del cual esta tesis aborda.

El descubrimiento de reglas de asociación ha sido uno de los tópicos muy importantes en la minería de datos, cuyo propósito es el poder encontrar relaciones de asociación entre conjuntos de elementos en una base de datos.

Dentro de las tareas de la minería de datos, encontraremos que la asociación forma parte esencial de ello, y es aquí donde el descubrimiento de reglas de asociación viene a ser el objeto de estudio.

Para poder llegar al descubrimiento de reglas de asociación, en la mayoría de los casos, se tiene que considerar el proceso de descubrimiento de conocimiento en bases de datos a través del análisis de la base de datos.

Otro campo interesante es la utilización del paradigma de Agentes dentro de la minería de datos, y a su vez la generación de reglas de asociación. Las razones por las que se ha introducido esta tecnología son varias, entre las que podemos mencionar y que también queremos tomar en cuenta, es el hecho de poder tener una aplicación distribuida con diversos agentes (*Sistema Multi-Agente* -SMA), que colaboren entre ellos para resolver una problemática para lo cual son diseñados.

Los SMA utilizan una arquitectura basada en servicios, lo que daría flexibilidad a nuestro sistema y así poder incluir agentes que realizan un servicio sobre un problema específico.

La idea de tener un SMA aplicado a reglas de asociación es el hecho de poder tener flexibilidad en nuestro sistema y extraer información (series de tiempo) sobre fuentes distribuidas.

1.2 Objetivos

En esta sección se presenta el objetivo general de la tesis y los objetivos específicos que permitan alcanzarlo.

1.2.1 Objetivo General

Desarrollar un sistema multiagente para la obtención de reglas de asociación mediante el análisis de series de tiempo en fuentes de datos distribuidas.

1.2.2 Objetivos Específicos

Para alcanzar el objetivo general del presente trabajo, se proponen los siguientes objetivos específicos:

1. Proponer, desarrollar y aplicar una metodología para generar reglas de asociación mediante el análisis de series de tiempo.
2. Desarrollar un sistema multiagente que permita la extracción de reglas de asociación a partir del análisis de series de tiempo.
3. Implementar el caso de estudio usando la plataforma de agentes CAPNET (Component Agent Platform on .NET).

1.3 Organización del documento

El contenido de la tesis se encuentra estructurado de la siguiente manera:

Capítulo 2 Se hace un tratado del estado de arte relacionado a sistemas que han incursionado en la generación de reglas de asociación a partir de series de tiempo, haciendo una descripción breve de los mismos y sus principales aportaciones.

Capítulo 3 Se presenta el marco teórico de aquellos conceptos que son base para el desarrollo de la presente tesis.

Capítulo 4 En éste capítulo se describe la metodología propuesta para la generación de reglas de asociación, incluyendo los conceptos básicos y un ejemplo de generación de reglas de asociación.

Capítulo 5 Se hace un tratado sobre el análisis y diseño del sistema multiagente, se presenta la arquitectura del sistema multiagente utilizada para la generación de reglas de asociación.

Capítulo 6 Se presenta el propotipo desarrollado para aplicar la metodología descrita en el capítulo anterior, describiendo los resultados obtenidos en el caso de estudio desarrollado.

Capítulo 7 Se presentan las conclusiones, contribuciones y el trabajo futuro de la presente tesis.

Capítulo 2

Estado de Arte

En el presente Capítulo se hace un tratado del estado de arte, donde se hace mención de algunos sistemas que hacen uso de la minería de datos para la generación de reglas de asociación y en otros casos aquellos que hacen uso del paradigma de sistemas multi-agente para la generación de las mismas.

2.1 Introducción a la minería de datos

La información generada por las empresas, centros de salud, departamentos gubernamentales, además del aumento de la información en la Internet que radica en un ambiente totalmente heterogéneo; se han desarrollado diversas tecnologías como la fusión de datos que además de recuperar, coleccionar e integrar datos, se enfoca en la obtención de nuevo conocimiento a partir de los datos fusionados.

La minería de datos está dedicada a la extracción automática de patrones desconocidos de datos dados, presenta diversas técnicas que han sido desarrolladas incluyendo la búsqueda de similaridad basada en patrones, análisis de cluster, clasificación basada en árboles de decisión y minería de reglas de asociación [Klusch, Lodi & Moro,2003].

Jianfeng Wu y Luicio Soibelman han trabajado en el campo de la fusión de datos y descubrimiento de conocimiento en la construcción de planificadores de datos aplicando técnicas de análisis de datos para encontrar conocimiento en base de datos en proyectos de planificación existentes. Su trabajo de investigación está enfocado a tres problemas de análisis para el descubrimiento de conocimiento. Primeramente en cómo planear y construir planificadores de proyectos existentes que fueron organizados en diferentes formatos y codificados en diferentes sistemas, permitiendo que las herramientas de fusión de datos agrupen y consoliden estas fuentes de datos para

la tarea de aprendizaje. Segundo, la fusión de planificadores de datos de proyectos históricos son representados de una manera conveniente para el proceso de descubrimiento de información de una manera eficaz. Y en tercer lugar, las técnicas de análisis de datos, las cuales fueron inicialmente desarrolladas por investigadores de las ciencias computacionales para identificar patrones y tendencias de datos transaccionales en bases de datos relacionales, necesitan ser adaptadas para permitir el descubrimiento de información mediante planificadores gráficos y orientados a la Red que tienen una lógica compleja y fuentes restringidas [Wu & Soibelman,2005].

En el principio estas técnicas se habían desarrollado para datos centralizados pero debido al crecimiento de la Internet y de la distribución de datos, estas técnicas han sido modificadas o desarrolladas para que puedan ser aplicados a datos distribuidos, contribuyendo al desarrollo de la minería de datos distribuida (*Distributed Data Mining* -DDM). La minería de datos distribuida pretende realizar el análisis de datos en sitios individuales, enviando los resultados parciales a otros sitios donde se pueda establecer un resultado global [Klusch, Lodi & Moro,Julio2003].

Como se mencionó en párrafos anteriores, la minería de datos tiene la tarea de obtener conocimiento de diversos repositorios de información, de manera análoga, así como el minero extrae oro de las rocas, la minería de datos tiene como propósito extraer conocimiento de grandes volúmenes de información.

Para algunos autores toman el término de minería de datos como sinónimo de descubrimiento de conocimiento en base de datos.

La minería de datos comprende una serie de técnicas, algoritmos y métodos cuyo fin es la explotación de grandes volúmenes de datos de cara al descubrimiento de información previamente desconocida y que pueda ser empleada como ayuda a la toma de decisiones.

El proceso de descubrimiento de información consiste de una secuencia iterativa de los siguientes pasos [Han & Kamber,2001]:

1. *Filtración de Datos.* Eliminar ruido e inconsistencia en los datos.
2. *Integración de Datos.* Multiples fuentes de datos pueden ser combinados.
3. *Selección de Datos.* Los datos relevantes para la tarea de análisis son recuperados desde la base de datos.
4. *Transformación de Datos.* Los datos son transformados o consolidados en formas apropiadas para la minería.
5. *Minería de Datos.* Es el proceso esencial donde métodos inteligentes son aplicados para extraer patrones de datos.

6. *Evaluación de Patrones.* Identificar patrones verdaderamente interesantes representando el conocimiento basado en algunas medidas de interés.

7. *Representación del Conocimiento.* Las técnicas de representación de conocimiento son usadas para presentar el conocimiento extraido al usuario.

Las técnicas de minería de datos son herramientas que facilitan el descubrimiento de la información, las más comunes son las siguientes [Hand et al,2001]:

- *Análisis de exploración de datos*: Como su nombre lo indica, su objetivo es explorar sin alguna idea clara de lo que se esta buscando. Típicamente estas técnicas son interactivas y visuales.

- *Modelo descriptivo*: El objetivo de esta técnica es describir todos los datos (o los procesos que generan los datos). Algunos ejemplos de estos modelos incluyen distribución probabilistica de los datos (*density estimation*), particionamiento de un espacio p-dimensional en grupos (*cluster analysis and segmentation*) y modelos que describen relaciones entre variables (*dependency modeling*).

- *Modelo predictivo (Clasificación y Regresión)*: Su propósito es construir un modelo que permita que el valor de una variable sea predecida de valores conocidos de otras variables. En la clasificación, la variable que está siendo predecida es categórica, mientras que en la regresión la variable es cuantitativa.

- *Descubrimiento de patrones y reglas*: Los tres tipos de tareas descritas anteriormente hacen referencia a la construcción del modelo. Exiten otras aplicaciones en la minería de datos que hacen referencia a la detección de patrones. Una de las tareas es encontrar combinaciones de elementos que ocurren frecuentemente en bases de datos transaccionales; este problema ha enfocado mucho la atención en la minería de datos, direccionandolos al uso de algoritmos basados en *reglas de asociación.*

- *Recuperación por contenido*: En este caso el usuario tiene un patrón de intereses y deseos para encontrar patrones similares en un conjunto de datos. Esta tarea es comunmente usada para un conjunto de imágenes y texto.

Un importante tópico en la minería de datos es relativo al descubrimiento de reglas de asociación. Una regla de asociación interesante, describe una relación interesante entre diferentes atributos. El problema de las reglas de asociación en la minería de datos, es identificar todas las reglas cuya confianza y soporte sean mayores al mínimo soporte y mínima confianza [Chan & Au,1997].

2.2 Sistemas que generan Reglas de Asociación

José Fernando Reyes Saldaña y Rodolfo García Flores, estudiantes de posgrado en Ingeniería de Sistemas de FIME-UANL, hacen referencia al proceso de descubrimiento de conocimiento en bases datos, considerando un caso de estudio, donde se extrajo un conjunto de reglas de asociación mediante el algoritmo conocido como Apriori. Este conjunto de reglas permite realizar el análisis de los patrones de compras de productos por parte de los clientes, y con esto obtener aplicaciones prácticas como son estrategias de compra y venta, de acomodo de productos diseño de promociones, entre otras. Aplicados en una empresa dedicada a la comercialización de productos químicos especializados. [Reyes & García,2005]

En un artículo publicado por Elena Luciv y B. Novikov de la Universidad de San Petersburgo, hacen referencia al descubrimiento de reglas de asociación en secuencias temporales, definiendo un secuencia temporal como un conjunto de valores reunidos en un cierto tiempo; ellos plantean una metodología para obtener dependencias entre patrones en secuencias temporales, permitiendo encontrar la relaciones causa-efecto entre los patrones temporales en una o más secuencias temporales, cuyos resultados pueden ser interpretados por un experto en el dominio o ser usados en un sistema basados en conocimiento. [Luciv & Novikov,2005]

Otro campo donde vemos el uso de reglas de asociación es en la detección de anomalías en trasbordadores espaciales. En [Yairi et al,2001] se propone un nuevo método de detección de anomalías para transbordadores espaciales basados en dos diferentes técnicas de minería de datos: reglas de asociación y clustering de patrones en series de tiempo. En este método, típicamente patrones temporales son extraídos de cada serie de tiempo de house-keeping data, los cuales han sido acumulados desde la primera fase de la operación inicial, luego entonces, las relaciones causa-efecto entre los patrones de diferentes series de tiempo son explorados y obtenidos en forma de reglas de asociación. El conjunto de reglas de asociación pueden ser reunidos como un modelo del trasbordador espacial y pueden ser usados para detectar si el comportamiento del sistema es normal o no. Este alcance tiene dos características notables, comparadas con los métodos de detección tradicionales. Primeramente éste requiere de un conocimiento a priori del sistema del trasbordador, obtenido directamente del house-keeping data. Por esta razón, este método puede ser aplicado a diferentes tipos de trasbordadores con un costo relativamente bajo. Segundo, éste modela el comportamiento del trasbordador a través de un conjunto de reglas de asociación, el cual es diferente en su representación o en las ecuaciones diferenciales o valores límite. Como resultado, éste espera detectar algunas pequeñas anomalías las cuales han sido revisados en los métodos convencionales. [Yairi et al,2001]

El Centro de Investigación de Almaden de IBM, desarrollaron un sistema de minería de datos llamado The Quest Data Mining System, este proyecto tenia el

propósito de desarrollar tecnología que permitiera nuevos enfoques en aplicaciones relativas a la toma de decisiones. El enfoque principal era incluir operaciones de minería de datos, que permitiera desarrollar aplicaciones de manera rápida, y que incluyera algoritmos escalables para su ejecución, teniendo las siguientes tareas: a) descubrir patrones en grandes bases de datos, en vez de verificar la existencia de un sólo patrón. b) tener una propiedad completa que garantice que todos los patrones del mismo tipo sean descubiertos y c) tener un alto performance y escalar linealmente en bases de datos de gran magnitud de la vida real [Agrawal & Arning et al,1996].

[Batyrshin et al,2004] proponen una metodología para generar reglas de asociación del tipo: *If Cond then A está asociado con B, (W)*, donde *Cond* es una restricción crisp en las series de tiempo, y la asociación es obtenido como coeficiente de correlación entre la restricción de las serie de tiempo A y B que describen algunos parámetros del sistema analizado. W es un significado de una regla dada por los valores del coeficiente de correlación.

En la minería de datos, el análisis de series de tiempo, como se mencionó líneas atrás, se espera encontrar asociaciones entre patrones existente en ellas, sobretodo aquellos en los cuales se presentan en intervalos de tiempo T, o también conocido como *windows* y un conjunto D contiene los posibles pares de *windows*.

En la minería de datos difuso (*fuzzy data mining)*, los conjuntos de elementos A y B, pueden denotar algunas propiedades difusas y las reglas de asociación pueden ser dadas por medio de tuplas definidas en dominios de atributos. El conjunto D contiene un conjunto de posibles tuplas y una propiedad difusa A define un confunto difuso sobre un conjunto de tuplas [De Cock et al,2005] [Dubois et al,2005].

Batyrshin y Sheremetov proponen una metodología para generar reglas de asociación para conjuntos difusos basados en percepciones difusas en patrones como un rápido incremento, un lento incremento, etc. Esta metodología fue aplicada a un sistema de análisis de indicadores de economía mexicana [Batyrshin & Sheremetov,2006].

La Universidad Politécnica de Hong Kong, desarrolló una técnica llamada F-APACS, que permite generar reglas de asociación difusas; esta técnica emplea terminos lingüísticos que permite la representación de regularidades y excepciones en base de datos. Esta representación hace que las reglas de asociación generadas sean mucho más manejables por el entendimiento humano.

F-APACS tiene la ventaja de que permite obtener reglas de asociación positivas y negativas. Una regla de asociación positiva nos dice que un registro tiene ciertos valores de atributos que también tendrá otro valor de atributo, mientras que una regla de asociación negativa nos dice que un registro tiene cierto valor de atributo que no tendrá otro valor en el atributo [Keith & Whai,1997].

2.3 Agentes y Plataformas de Agentes

2.3.1 Agentes de Software

La tecnología de agentes es un área de investigación de interés en nuestros días, desde la aparición de la Inteligencia Artificial Distribuida, cuya finalidad es desarrollar ambientes distribuidos inteligentes, llamados *agentes*, que interactúan mediante la cooperación, la competición, y la negociación.

El concepto de agentes es uno de los más importantes en los años 90, tanto en la Inteligencia Artificial Distribuida como en las Ciencias de la Informática y Computación. Se han desarrollado aplicaciones en campos tan variados como la administración en las telecomunicaciones, el control de tráfico aéreo, la minería de datos, la recuperación de la información, el comercio electrónico, los asistentes personales digitales, las librerías digitales, el control de procesos, las bases de datos inteligentes y la educación.

Nwana [Nwana & Ndumu,1998] define a un agente como: *"Un componente de software y/o hardware, capaz de realizar una tarea en favor de un usuario"*.

Shoham [Shoham,1993] define a un agente de software como: *"Una entidad de software la cual funciona continuamente y autónomamente en un ambiente en particular, a menudo habitada por otros agentes y procesos"*.

Existen muchas definiciones del concepto de Agente.

[Wooldrigde & Jeannings,1994]Wooldridge define la noción débil y la noción fuerte de agente como se indica a continuación:

Noción débil de agente:

El término de agente es usado para denotar aquel hardware o sistema computacional que tiene las siguientes características:

- *Autonomía:* los agentes operan sin la intervención directa de los seres humanos, tienen estados internos, y tienen alguna clase de control sobre sus acciones.
- *Capacidad Social:* los agentes interactúan con otros agentes (posiblemente seres humanos), mediante un tipo de lenguaje de comunicación entre agentes.
- *Reactividad:* los agentes perciben su ambiente (el cual puede ser el mundo físico, un usuario a través de una interfaz gráfica, Internet, o todas combinadas) y responden de una manera oportuna frente a los cambios que ocurren en él.

- *Pro - actividad:* los agentes no simplemente actúan como respuesta a su ambiente, son capaces de exhibir un comportamiento dirigido a metas mediante la toma de iniciativas propias.

Noción fuerte de agente:

Un agente es considerado como un sistema computacional que, además de tener las características antes mencionadas, tienen atributos (estados mentales) propios de los seres humanos como *conocimiento, creencias, intenciones y deseos*, entre otros.

Shoham [Shoham,1993] define a un agente como: *"Una entidad de software la cual funciona continuamente y autónomamente en un ambiente en particular, a menudo habitada por otros agentes y procesos"*.

Cuando hablamos de agentes de software, existen tres dimensiones para poder medir su capacidad, las cuales son: Agencia, inteligencia y movilidad. Agencia es el grado de autonomía y autoridad establecida en el agente; inteligencia es el grado de razonamiento y conducta aprendida para la solución de problemas; movilidad es la capacidad que tienen los agentes para moverse a través de la red.

Wooldrigde [Wooldrigde & Jeannings,1994] menciona otros atributos en el contexto de agencia. Por ejemplo:

- *Movilidad* es la habilidad de un agente de moverse alrededor de una red electrónica.
- *Veracidad* es la suposición de que un agente no comunicará intencionalmente información falsa.
- *Benevolencia* es la suposición de que los agentes no tienen metas conflictivas, y que cada agente siempre tratará de hacer lo que se le pide que haga.
- *Racionalidad* es la suposición de que un agente actuará en orden de conseguir sus metas, y no actuará de tal forma como para prevenir que sus metas sean conseguidas, al menos hasta el punto que sus creencias lo permitan.

2.3.2 Sistemas multiagentes

La tecnología de agentes ha despertado el desarrollo de aplicaciones basadas en ellos, llegándose no solo a desarrollar metodologías para el desarrollo de sistemas basados en agentes, sino también se ha introducido esta tecnología en diversas áreas de investigación, inclusive en la fusión de datos y descubrimiento de información.

Matthias Klusch y su equipo de investigadores se han enfocado en el área de minería de datos distribuida, aplicando la tecnología de agentes en sus proyectos.

Ellos pretenden desarrollar un sistema que contiene una sociedad de agentes, con el fin de que estos agentes cooperen entre sí para alcanzar una meta en común extrayendo información de diversas bases de datos, donde cada agente evalúa la ventaja y el riesgo de hacer la tarea de minería en los datos [Klusch, Lodi & Moro,Julio2003].

En el Departamento de Ciencias Computaciones de la Universidad de Aberdeen en el Reino Unido, aplican la tecnología de agentes en la minería de datos sobre bases de datos distribuidas, pretendiendo diseñar un agente de minería de datos abierto y flexible, teniendo un grupo de agentes que puedan cooperar entre sí para descubrir conocimiento en fuentes de datos distribuidas. Ellos exploran el uso de un nuevo lenguaje de agentes, Agent -k basado en la programación lógica inductiva (*Inductive Logic Programming* -ILP), permitiendo así no solo la integración sino la inducción del conocimiento [Winton & Pete,1995].

Se han desarrollado metodologías para el desarrollo de aplicaciones basados en Agentes, Zili Zhang se enfoca específicamente al desarrollo de un Framework Híbrido basado en agentes para integrar diversas técnicas de minería en bases de datos. Este framework está basado en la metodología GAIA desarrollada por Wooldridge para crear sistemas basados en agentes. Lo relevante de este framework es permitir a nuevas técnicas KDD puedan integrarse al sistema y que las técnicas obsoletas puedan ser eliminadas del sistema dinámicamente, además de que las técnicas KDD basadas en agentes puedan interactuar en tiempo de ejecución dentro del framework; para aquellos sistemas no basados en agentes, sus interacciones deben ser determinadas en el tiempo de diseño, siendo el objetivo de este framework el proveer una plataforma robusta y flexible para la minería de datos [Zhang,2003].

2.3.3 Ambientes de desarrollo

Para la construcción de aplicaciones basadas en Agentes se han desarrollado herramientas donde la primera generación de éstas ofrece a los usuarios las herramientas para el desarrollo de agentes y sistemas multiagente, sin embargo no se apegaban a algún estándar. A continuación mencionaremos brevemente algunos sistemas de este tipo:

2.3.3.1 Agent Builder

[Agent Builder,2002]Es un producto comercial producido por Reticular System, Inc. Este proporciona una herramienta para la construcción de agentes inteligentes, proporcionando interfaces gráficas para el diseño y desarrollo de sistemas multiagentes. Está basado en Java y su lenguaje de comunicación es KQML.

Proporciona herramientas para el análisis del dominio del problema, así también

de herramientas para definir la agencia (colección de agentes inteligentes), integra y usa librerías de Java, C y C++.

2.3.3.2 Madkit

Madkit, es una plataforma para desarrollo de sistemas multi-agente, la cual esta basada en Java. Fue desarrollado por Oliver Gutknecht y Jacques Ferber en el LIRMM (Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier), en una investigación de laboratorio en Francia. Contrariamente a otras plataformas, MadKit es principalmente una máquina en línea de SMA(Multi Agent System, MAS), usando un agente micro-kernel [Ferber & Gutknecht,1998].

2.3.3.3 JATLite

JATLite (*Agent Template Lite*), la cual fue desarrollada por la Universidad de Stanford, provee un paquete de programas escritos en lenguaje Java que permite a los usuarios crear rápidamente software de agentes que se comunican robustamente sobre el Internet. JATLite proporciona una infraestructura básica en la que los agentes se registran por un Agente Planificador-Routeador de Mensajes, que usa un nombre y contraseña, conecta / desconecta del Internet, envía y recibe mensajes, transfiere archivos, e invoca otros programas o acciones en las varias computadoras donde ellos están corriendo. La parte interesante de JATLite es el AMR (*Agent Message Router*) el cual permite a los agentes recuperar, migrar y ser applets. Un tradicional ANS proporciona solamente la dirección de cuando un agente requiere a otro agente. Cada agente es responsable de guardar las direcciones IP de todos los agentes que ellos les corresponden [Heecheol et al,2000].

La mayor desventaja de esta situación radicaba en el hecho que los sistemas desarrollados en entornos diferentes no eran compatibles entre sí. La Federación para los Agentes Físicos Inteligentes (FIPA), es una organización encargada de estandarizar las tecnologías basadas en agentes, la cual ha llegado a ser un estándar muy importante para el desarrollo de plataformas de agentes influyendo en la aparición de la segunda generación de ambientes de desarrollo, los cuales son compatibles con los estándares de FIPA.

Los estándares de FIPA definen un modelo de referencia común para una plataforma de agentes (*Agent Platform -AP*) como un conjunto de cuatro componentes donde cada uno representa las capacidades lógicas o servicios que pueden ser combinados en cada implementación concreta del AP: Agentes, Directory Facilitator (DF), Agent Management System (AMS) y Message Transport System (MTS). El AMS y el DF son agentes dedicados a dar soporte en la administración de otros agentes, mientras que el MTS proporciona de un servicio de entrega de mensajes (Ver Fig.2.1). A continuación presentamos algunos sistemas de mayor uso:

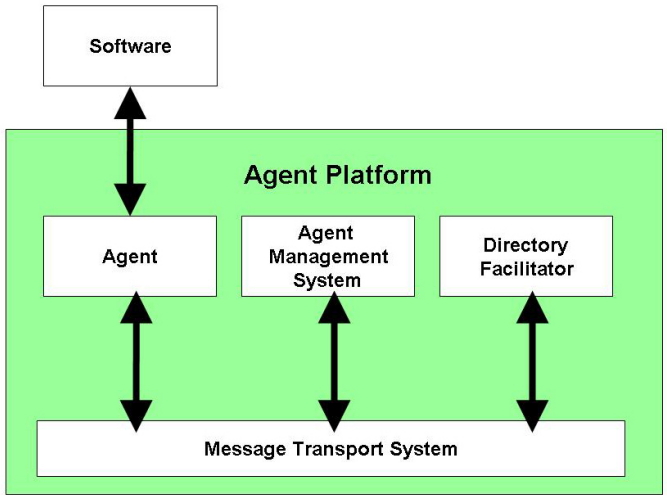


Figura 2.1: *Modelo de referencia de FIPA*

2.3.3.4 Jade

JADE (Java Agent Development Framework) es un framework implementado en Java, el cual simplifica el desarrollo de sistemas multiagentes, de acuerdo a las especificaciones que FIPA (Foundation for Intelligent Physical Agents) ha establecido. Proporciona un conjunto de interfaces para el desarrollo de agentes implementados en Java. JADE usa el lenguaje de comunicación de agentes de FIPA, utilizando una combinación de socket, RMI y CORBA.

2.3.3.5 Zeus

Zeus, es una herramienta desarrollada por Agent Research Programme de la British Telecom Intelligent System Research Laboratory, es un framework para el desarrollo de sistemas de agentes colaborativos. Zeus fue construido en Java a causa de la portabilidad y el soporte de multithread; su metodología usa una descomposición de cuatro partes para el desarrollo de agentes: análisis, diseño y soporte [Zeus,1999].

Existen tres grupos de clases en Zeus, una librería de los componentes de un agente, un conjunto de herramientas visuales y el software para construir agentes. Un agente Zeus está compuesto por tres capas: una capa de definición, una capa organizacional y la capa de coordinación. La capa de definición representa las capacidades del agente DBI, la capa de organización define las relaciones con otros agentes, la capa de coordinación modela cada agente como una entidad.

2.3.3.6 CAPNET

CAPNET (*Component Agent Platform based on .NET*), es una plataforma para el desarrollo, despliegue y administración de Sistemas Multi-Agente compatible con especificaciones internacionalmente adoptadas; desarrollado por el Instituto Mexicano del Petróleo (IMP), basado en las especificaciones de FIPA y escrito completamente en el framework .NET de Microsoft, además utiliza estándares ampliamente adoptados por la industria tales como: Infraestructura .NET, XML, RDF, Web Services y obviamente FIPA.

El propósito de CAPNET es permitir que los desarrolladores puedan crear e integrar aplicaciones distribuidas basadas en la tecnología de agentes, además de contar con la posibilidad de interoperar con aplicaciones desarrolladas en otras plataformas de agentes, también brinda un conjunto de: Servicios, Herramientas, Ambiente de desarrollo y Entorno de ejecución de SMA (Ver Figura 2.2).

Figura 2.2: *Estructura de CAPNET*

Arquitectura de CAPNET

CAPNET cuenta con una arquitectura de 4 niveles: La aplicación de agentes, el administrador de agentes y el directorio de servicios, el nivel de seguridad y técnicas de conectividad, y el sistema de transporte de mensajes, como se muestra en la Figura 2.3.

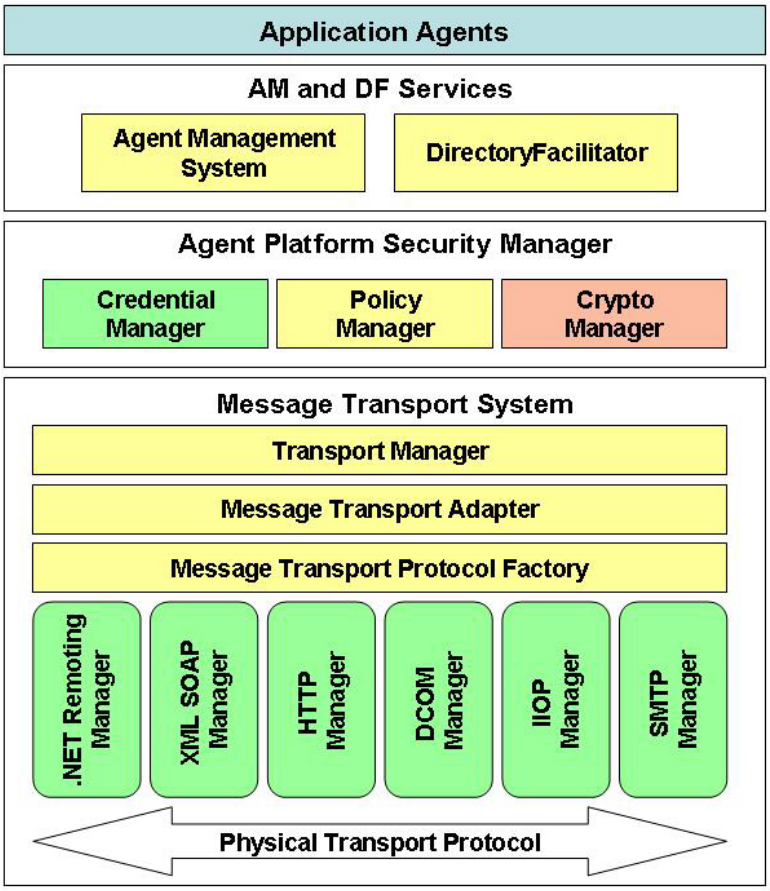


Figura 2.3: *Arquitectura de CAPNET*

La capa de aplicación de agentes es aquella en la cual los agentes interactúan entre sí y con la plataforma; los servicios de administración de la plataforma, específicamente el AMS y el DF, se encuentra en la siguiente capa de aplicación de agentes, puesto que ahí se administra el ciclo de vida de los agentes y del directorio de servicios proporcionados por cada uno de ellos; la capa de administración de seguridad, es la responsable de mantener las políticas de seguridad de la plataforma, la autentificación y actividades en tiempo de ejecución tales como la comunicación y proveer seguridad en el nivel de transporte. Por último se encuentra el servicio de transporte, en el cual la entrega y la recepción de los mensajes vienen a ser un punto muy importante dentro de la plataforma CAPNET, ya que la comunicación está basada en sistemas de mensajería débilmente acoplados [Contreras et al,2004].

Capítulo 3

Marco Teórico

En este Capítulo se presenta un tratado de conceptos teóricos, considerando aquellos temas que son de gran importancia para el entendimiento y desarrollo de la presente tesis.

3.1 Métodos básicos de clasificación y cluster análisis

3.1.1 Clasificación

Hablar de clasificación implica hablar de un tópico que ha sido empleado en diversas disciplinas, como expresa Michie et al [Michie et al,1994], que la clasificación podría cubrir cualquier contexto en alguna decisión o prevención hecha en base a la información actualmente disponible y el proceso de clasificación es un método formal para emitir juicios repetidamente en nuevas situaciones.

De acuerdo al diccionario de la Real Academia Española, la clasificación es el acto y acción de clasificar, entendiéndose clasificar al hecho de ordenar o disponer por clases; en otras palabras, es aquel procedimiento para construir agrupaciones o categorías en base a atributos o relaciones comunes. La clasificación es a menudo un medio práctico y útil de organizar pequeñas cantidades de datos [Chou,1977].

3.1.2 Cluster Análisis

Cluster Análisis es un conjunto de técnicas utilizadas para clasificar los objetos o casos en grupos homogéneos llamados conglomerados (*clusters*) con respecto a algún

criterio de selección predeterminado. Los objetos dentro de cada grupo (conglomerado), son similares entre sí (alta homogeneidad interna) y diferentes a los objetos de los otros conglomerados o clusters (alta heterogeneidad externa). Es decir, que si la clasificación hecha es óptima, los objetos dentro de cada cluster estarán cercanos unos de otros y los cluster diferentes estarán muy apartados. Por ello, es también conocido como análisis de clasificación o taxonomía numérica.

El análisis por cluster, como también es conocido, es una técnica que agrupa a los elementos de una muestra en grupos llamados conglomerados, siendo cada conglomerado lo más homogéneo que sea posible y a la vez los conglomerados sean muy distintos entre si.

Existen varios tipos de análisis de cluster, pero Klijin hace referencia a una clasificación conocida como Taxonomía Numérica, la cual es una clasificación de tipo jerárquico y que tiene su origen en la sistemática del reino animal y vegetal. La clasificación jerárquica parte de un conjunto Ω cuyos elementos debes ser clasificados. Se trata de obtener sucesiones particiones ("*clusterings*"), organizadas en diferentes niveles jerárquicos, estando cada partición formada por clases disjuntas ("*clusters*"). Los elementos de una misma clase deben ser razonablemente homogéneos.

Características importantes de una clasificación (jerárquica).

Relaciones entre los objetos

Ellas se establecen calculando una matriz de similaridades o de disimilaridades que informen sobre las analogías o diferencias entre unos y otros, sobre la base de las características cualitativas elegidas. Cuando se utilizan variables cuantitativas se trabaja con una matriz de correlaciones (correlación de Pearson) o de distancias (euclídea, de Minkowski, de Mahalanobis, etc).

Tipos de clasificación

1. Aglomerativa -Divisiva: En una clasificación aglomerativa se parte inicialmente de los objetos, que se van progresivamente fusionando para formar particiones sucesivas; en una clasificación divisiva se parte del conjunto total Ω que se subdivide progresivamente hasta alcanzar un grado aceptable de subdivisión.

2. Jerárquica - No jerárquica: En una clasificación no jerárquica se forman grupos homogéneos sin establecer relaciones entre los grupos; en una clasificación jerárquica los grupos se van fusionando progresivamente, mientras decrece la homogeneidad entre los grupos, cada vez más amplios, que se van formando. Una clasificación jerárquica es en general aglomerativa.

3. Monotética - Politética: Una clasificación monotética está basada en una característica única que sea muy relevante. Es divisiva, pues los objetos se clasifican

en los que tienen la característica y los que no la tienen. Puede dar lugar a clasificaciones poco adecuadas, dada la dificultad de obtener grupos bastante homogéneos y naturales. Una clasificación politética está basada en un número grande de características, y no exige que todos los elementos de una clase posean todas las características, sino un número suficientemente grande para poder justificar analogías entre miembros de una misma clase. Este tipo de clasificación es aglomerativo [Klijin,2001].

3.2 Medidas de Asociación

La razón por la cual se estudian las variables es para ver si existe relación entre ellas, pudiendo "predecir" (en caso de haberla) valores de una a partir de la otra. Una forma de detectar la posible relación entre las variables es gráficamente, y el gráfico utilizado es conocido como *diagrama de dispersión* o *nube de puntos*. Otra forma es a través de medidas numéricas tales como la covarianza o el *coeficiente de correlación de Pearson*.

En esta sección se define el coeficiente de correlación, algunas propiedades y también se considera la matriz de correlación.

El coeficiente de correlación: Es una medida del grado de relación existente entre dos variables. Se define como:

$$\rho_{x,y} = \frac{\sigma_{x,y}}{\sigma_x \sigma_y} \tag{3.1}$$

Algunas observaciones y propiedades:

1. Su signo es determinado por la covarianza, indicando si la asociación es positiva o negativa, cuando la covarianza tiene el valor de 0, de la misma manera su valor será 0. (Ausencia de asociación lineal).

2. $-1 \leq \rho_{x,y} \leq -1$ Valores próximos a -1 indican fuerte asociación lineal negativa, valores próximos a 1 indican fuerte asociación lineal positiva, y valores próximos a 0 indican ausencia de asociación lineal.

3. No se debe de interpretar el coeficiente sin haber visto previamente el diagrama de dispersión.

4. Un coeficiente de correlación alto (en valor absoluto) indica que las variables toman valores relacionados entre sí entre los elementos observados, pero no permite concluir la existencia de ninguna relación de casualidad entre las variables. Por ejemplo: suponiendo que se estudian las siguientes variables x =número de matrimonios mensuales (en una ciudad) y y =temperatura del mes, obteniéndose un coeficiente de correlación de 0.7 Esto significa que en efecto, suele haber mas matrimonios a medida que mejoran las temperaturas, pero no implica que un aumento de matrimonios aumente la temperatura del mes, ni que una ola de calor cause una avalancha de matrimonios.

El análisis de correlación es la herramienta estadística que se usa para describir el grado en el que una variable está linealmente relacionada con otra. Los estadísticos han desarrollado dos medidas para describir la correlación entre dos variables: el *coeficiente de determinación* y el *coeficiente de correlación* [Levin & Rubin,2004].

La matriz de coeficientes de correlación R, es aquella matriz simétrica que en la posición (i,j) tiene el coeficiente de correlación $\rho_{x,y}$, la cual nos indica el grado de relación existente entre las variables.

3.3 Métodos de agrupación

Existen dos tipos básicos y se distinguen por su naturaleza jerárquica o no jerárquica.

Los métodos más eficientes de agrupación jerárquica se conocen como *métodos de agrupación de un solo enlace.*

Método del vecino más cercano.

Un ejemplo de un método de agrupación de un sólo enlace es el método del vecino más cercano. En éste se aplican los siguientes pasos [Johnson,1998]:

1. Empiece con N agrupamientos, en donde cada uno de ellos contiene exactamente un punto dado.

2. Enlace los dos puntos más cercanos según una de las tres medidas seleccionadas de la distancia.

3. Defina la desemejanza entre este nuevo agrupamiento y cualquier otro punto como la distancia mínima entre los dos puntos del agrupamiento y este punto.

4. Continúe combinando los agrupamientos que sean los más cercanos entre sí de modo que, en cada etapa, la cantidad de agrupamientos se reduzca en uno y la desemejanza entre cualesquiera dos de estos siempre se defina como la distancia entre sus miembros más cercanos.

De este modo, el método del vecino más cercano se inicia con N agrupamientos, en donde cada uno de estos contiene una observación y continúa combinando los puntos y agrupamientos hasta que todas las observaciones están dentro de un agrupamiento. Es evidente que el número apropiado de agrupamientos se encuentra en algún lugar entre el principio de este proceso y su final.

3.4 Minería de datos

3.4.1 Definición de Minería de Datos

El término de minería de datos o "*Data Mining*" en inglés, ha sido propuesto en las últimas décadas, el cual hace referencia a la extracción de conocimiento a partir de grandes volúmenes de información, obteniendo información útil como esto: "cuando un cliente compra leche, el cliente tambien comprará pan" y "clientes les gusta comprar productos para el sol".

Hablando estrictamente [Zhang & Zhang,2002]:

La minería de datos es un proceso de descubrimiento de información valiosa de grandes cantidades de datos almacenados en bases de datos, datawarehouses o algun otro repositorio de información.

Esta información valiosa puede ser patrones, asociaciones, cambios, anomalías y estructuras significantes, esto es, que la minería de datos pretende extraer conocimiento potencialmente útil a partir de datos.

La minería de datos ha sido popularmente tomado como sinónimo de KDD, aunque algunos investigadores ven a la minería de datos como una parte esencial del descubrimiento del conocimiento. Algunas de las aplicaciones son las siguientes:

- Detección de fraudes: Identificar transacciones fraudulentas.
- Aprobación de préstamos: Decidir que clientes son acreedores de un préstamo.
- Análisis de inversión: Predice un protafolio o una inversión.

- Tratos de Portafolio: Trata de un portafolio de instrumentos financieros maximizando regresos y minimizando riesgos.
- Mercadotécnia y análisis de venta de datos: Identificar clientes potenciales, establecer la efectividad en campañas de venta.
- Análisis del proceso de manufacturación: Identificar las causas de problemas en la manufacturación.
- Análisis de resultados de experimentos: Resume los resultados y predice modelos.
- Análisis de datos científicos.
- Agentes inteligentes y navegación WWW.

3.4.2 Tareas de la Minería de Datos

De manera general las tareas de la minería de datos pueden ser clasificadas en dos categorías: *Minería de datos Descriptiva* y la *Minería de Datos Predictiva*.

Un sistema de minería de datos puede estar acompañado de uno o más de las siguientes tareas [Zhang & Zhang,2002]:

1. *Descripción de clases*. Una descripción de clases provee una descripción precisa y clara de una colección de datos y los distingue uno del otro.
2. *Asociación*. La asociación es el descubrimiento de relaciónes de asociación o correlación en un conjunto de elementos.
3. *Clasificación*. La clasificación analiza un conjunto de datos y construye un modelo para cada clase basado en las caracteristicas de los datos. Un árbol de decisión o un conjunto de reglas de clasificación son generados en el proceso de clasificación, el cual es usado para la mejor comprensión de cada clase en la base de datos y para la clasificación futura de ellos.
4. *Predicción*. Esta tarea de la minería de datos, predice los valores posibles en ciertos datos, o el valor de la distribución de ciertos atributos en un conjunto de objetos. Esto involucra encontrar el conjunto de atributos relevantes para el atributo de interés (mediante análisis estadístico) y predecir el valor de la distribución basado en el conjunto de datos similares para los datos seleccionados. El análisis de regresión, el modelo lineal generalizado, el análisis de correlación

y árboles de decisión han sido herramientas útiles en la predicción. Los algoritmos genéticos y modelos de redes neuronales también han sido popularmente usados.

5. *Clustering.* El análisis de clustering o conglomerados, identifica clusters en los datos, donde un cluster es una colección de datos de objetos que son similares el uno al otro. Similarmente pueden ser expresados por funciones de distancia especificados por el usuario o por un experto.

6. *Análisis de Series de Tiempo.* El análisis de series de tiempo, analiza un conjunto grande de series de tiempo de datos para determinar cierta regularidad y caracteristicas interesantes, incluyendo búsquedas de secuencias o patrones, cambios y variaciones.

3.5 Técnicas de minería de Datos

La minería de datos, como se mencionó en el apartado anterior, no es simplemente una técnica, ya que cualquier método que ayude a obtener información es útil.

Cada método tiene un propósito diferente, presentando de manera individual sus propias ventajas y desventajas. Sin embargo, la mayoría de los métodos usandos en la minería de datos pueden ser clasificados en los siguientes grupos [Goebel & Gruenwald,1999].

1. *Métodos estadísticos*: Estos se han enfocado principalmente en probar hipótesis preconcebidas y colocarlos en modelos de datos. Estos métodos serán usados generalmente por estadísticos, por esta razón la intervención humana es requerida para la generación de hipótesis y modelos.

2. *Razonamiento basado en casos*: Case-Based Reasoning (CBR) es una tecnología que intenta resolver un problema dado haciendo uso directo de experiencias y soluciones del pasado. Un caso es normalmente un problema específico que se ha encontrado previamente y se ha resuelto. Dado un nuevo problema particular, el razonamiento basado en casos examina el conjunto de casos almacenados localizando uno silimiar al problema dado. Si los casos similares existen, su solución se aplica al nuevo problema, y el problema es agregado a la base del caso para su uso futuro.

3. *Redes neuronales*: Neural networks (NN) es una clase de sistemas que modelan al cerebro humano. Así como el cerebro humano consiste en millones de neuronas interconectadas por la sinopsis, las redes neuronales se forman de grandes números de neuronas simuladas conectados entre ellas como las neuronas del cerebro son conectadas. Similar al cerebro humano, la fuerza de las interconexiones de la neurona pueden cambiar (o son cambiados por el algoritmo de aprendizaje) en respuesta a un estímulo presentado u obtenido del exterior que permite que la red "aprenda".

4. *Árboles de Decisión*: Un árbol de decisión es un árbol donde cada nodo noterminal representa una prueba o decisión sobre los datos de elementos considerados. Dependiendo del resultado de la prueba, uno escoge una cierta rama. Para clasificar un elemento de datos particular, iniciamos con el nodo raíz siguiendo por las ramificaciones hasta llegar al nodo terminal (u hoja). Cuando un nodo terminal es alcanzado, una decisión es hecha. Los árboles de decisión pueden ser interpretados como una forma especial de un conjunto de regla, caracterizado por su organización jerárquica de reglas.

5. *Inducción de reglas*: Las Reglas presentan una correlación estadística entre la ocurrencia de ciertos atributos en un elemento de datos, o entre ciertos elementos de datos en un conjunto de datos. La forma general de una regla de la asociación es $Xl\wedge, ...,\wedge Xn \Longrightarrow Y[C, S]$, donde los atributos $X1, ..., Xn$ predicen Y con una confianza C y una significancia S.

6. *Redes de Creencia Bayesiana*: Bayesian Belief Networs (BBN), son representaciones gráficas de distribuciones de probabilidad, derivadas del conteo de ocurrencias en el conjunto de elementos de datos. Específicamente, un BBN es un grafo dirigido acíclico, donde los nodos representan variables del atributo y las aristas representan dependencias probabilisticas entre las variables del atributo. Asociado con cada nodo están las distribuciones probabilisticas condicionadas que describen las relaciones entre el nodo y sus padres.

7. *Algoritmos genéticos / Programación Evolutiva*: Los algoritmos genéticos y la programación evolutiva son algoritmos estratégicos de optimización que están inspirados por los principios observados en la evolución natural. A partir de una colección de soluciones de problemas potenciales que compiten entre ellos, se seleccionan las mejores soluciones y combinaciones entre ellas, esperando que la bondad global del conjunto solución llegará a ser mejor y mejor, similar al proceso de evolución de una población de organismos. Los algoritmos genéticos y la programación evolutiva son usandos en la minería de datos para formular hipótesis sobre las dependencias entre las variables, como lo hacen las reglas de asociación o algún otro formalismo.

8. *Conjutos Difusos*: Los conjuntos difusos forman una metodología clave para representar y procesar incertidumbre. La incertidumbre aparece en muchas formas en las bases de datos en la actualidad: imprecision, inconsistencia, vaguedad, etc. Los conjuntos difusos manejan la incertidumbre en un esfuerzo por hacer un sistema complejo manejable. Así como los conjuntos difusos constituyen un poderoso alcance para tratar no sólo con datos incompletos, ruidosos e imprecisos, sino que también pueda ser útil en el desarrollo de modelos inciertos de los datos que proveen una ejecución más inteligente y más suave que los sistemas tradicionales. Desde que los sistemas difusos pueden tolerar incertidumbre y subsecuentemente utilizan un lenguaje como vaguedad para datos ligeros, ellos pueden ofrecer modelos robustos, tolerantes al ruido o predicciones en situaciones donde la entrada precisa no está disponible o es demasiado costosa.

9. *Conjuntos robustos o aproximados*: Un conjunto robusto es definido por el límite bajo y superior de un conjunto. Cada miembro del límite inferior es un cierto miembro de un conjunto. Cada no-miembro del límite superior es un cierto no-miembro de un conjunto. El límite superior de un conjunto robusto es la unión entre el límite inferior y el límite de una región llamado frontera. Un miembro de región frontera es posiblemente (pero no ciertamente) un miembro del conjunto. Por consiguiente, los conjuntos robustos pueden ser vistos como conjuntos difusos con un tercer valor de la función miembro (sí, no, quizá). Semejante a los conjuntos difusos, los conjuntos robustos son un concepto matemático que maneja incertidumbre en los datos. Similar a los conjuntos difusos, los conjutos robustos se usan raramente como una solución stand-alone, ellos son combinados normalmente con otros métodos como la inducción de reglas, clasificación o métodos de clustering.

Capítulo 4

Reglas basadas en correlación

El presente Capítulo tiene el propósito el definir los conceptos básicos de reglas de asociación, así como presentar la metodología usada para la generación de las mismas a partir del análisis de series de tiempo usando el paradigma de agentes.

4.1 Introducción

Debido al crecimiento de información manejada en diversos centros de trabajo, tanto de investigación como gubernamentales, esta información se ha almacenado en bases de datos o repositorios electrónicos, donde la mayoría de estos datos se interpretan como un conjunto de transacciones.

Dado un conjunto I de objetos a los cuales se les llama *items*, una transacción es un subconjunto no vacío de I.

La extracción de reglas de asociación es una técnica en la minería de datos, la cual es comunmente aplicada para el descubrimiento de patrones en sistemas de aprendizaje no supervisado.

Una *regla de asociación* es una implicación del tipo $A \Rightarrow C$, donde $A, C \in I$, y además se verifica que $A \cap C = 0$. Teniendo como significado de esta regla en el que *"Si en una transacción aparece un conjunto de items A, entonces también aparece el conjunto de items C"*

A continuación abordaremos en detalle el concepto de reglas de asociación.

4.2 Reglas de Asociación, definiciones

Una *regla de asociación* está formada de dos componentes [Reyes & García,2005]: la premisa y la conclusión. Las reglas generalmente se describen con una flecha apuntando hacia la conclusión, por ejemplo: $\{0041\} \longrightarrow \{3495\}$; una regla de asociación indica una afinidad entre la premisa y la conclusión, generalmente es acompañada por estadísticos basados en frecuencia que describen esta relación.

Una de las principales características de las reglas de asociación es la presencia de *incertidumbre*, es decir, no siempre son exactas. Tenemos dos medidas para determinar el grado de cumplimiento y el interés de una regla, llamadas *confianza (confidence)* y *soporte (supp)*.

La *confianza* se define como el porcentaje de transacciones que contienen el conjunto A, que también contienen a C. El *soporte* es el porcentaje de transacciones que contiene simultáneamente A y C; se dice que una regla de asociación es fuerte cuando su confianza y soporte son mayores que dos umbrales definidos por el usuario: *minconf* y *minsop*.

Formalmente, una *regla de asociación* es definida de la siguiente forma:

Considere el siguiente conjunto $I = \{i_1, i_2, ..., i_m\}$, representando un conjunto de elementos(*items*); $A_i = v$ es un elemento donde precisamente v es el valor del atributo A_i, en una relación $R(A_1, ..., A_n)$.

X es un conjunto de elementos(*itemset*) si este es un subcojunto de I.

Un *itemset* X en una base de datos D, tiene un soporte denotado como $supp(X)$. Siendo el número de transacciones de D conteniendo a X.

O bien:

$$supp(X) = |X(t)|/|D|$$

donde $X(t) = \{t \in D \mid t \text{ contiene } X\}$.

Un *itemset* X en una base de datos D, es llamado un elemento frecuente si este soporta el mismo o mayor soporte(*minsupp*) dado por el usuario.

Una *regla de asociación* es la implicación $X \longrightarrow Y$, donde los itemsets X y Y no se intersectan.

Cada regla de asociación tiene dos medidas de calidad, el soporte(*support*) y la confianza(*confidence*), definidas como:

El soporte de una regla $X \longrightarrow Y$ es el soporte de $X \cup Y$; y la confianza de una regla $X \longrightarrow Y$ es $conf(X \longrightarrow Y)$ como el radio $|(X \cup Y)(t)|/|X(t)|$, o $supp(X \cup Y)/supp(X)$.

En otras palabras el soporte es igual a la frecuencia de ocurrencia de patrones, mientras que la confianza es la fortaleza de la implicación [Zhang & Zhang,2002].

Debido a que las reglas de asociación ha sido objeto de estudio en los últimos años, en [Botía et al,2005] se hace mención de las principales líneas de investigación sobre este ámbito y son las siguientes:

- *Algoritmos.* La búsqueda de reglas de asociación fuertes, implica describir conjuntos de items "frecuentes", es decir, aquellos cuyo soporte (porcentaje de transacciones en los que aparecen) sea mayor a *minsop*; utilizar los conjuntos de items frecuentes para generar reglas de asociación fuertes. Esto implica un esfuerzo hablando computacionalmente, debido a que las bases de datos suelen contener un gran número de transacciones, y también a que el número de conjuntos de items es exponencial con respecto al número de items. Las ventajas e inconvenientes de los algoritmos dependen de las características de los datos y del problema, tales como el número de items involucrados y el tamaño del conjunto de transacciones.

- *Reglas de Asociación cuantitativas y difusas.* En bases de datos relacionales, las reglas de asociación ligan la presencia de valores de atributos (*items*) en tuplas (*transacciones*). En este contexto, la aparición de atributos cuantitativos supone un doble problema. Por una parte, el gran número de valores que puede tomar un atributo cuantitativo hace que los conjuntos de items difícilmente puedan ser frecuentes. Por otro lado, las posibles reglas tendrían un contenido semántico pobre, pensemos en la regla $Peso = 85.3 \Longrightarrow Altura = 1.75$, por ejemplo. Un problema adicional es que la complejidad computacional del problema aumenta de manera importante. Se plantea un primer algoritmo, que basa la solución del problema en la división del dominio de atributos cuantitativos en intervalos. Esta técnica plantea a su vez problemas, derivados de la semántica de los intervalos y la sensibilidad del soporte de los mismos a pequeños cambios de las fronteras. Para solucionarlo se propone un nuevo enfoque basado en el uso de etiquetas linguísticas, definidas mediante conjuntos difusos sobre el dominio de los atributos cuantitativos. Las reglas que relacionan etiquetas linguísticas de este tipo reciben el nombre de reglas de asociación difusas.

- *Reglas de Asociación generalizadas.* En ocasiones, las bases de datos contienen información acerca de distintos niveles de abstracción de los datos, y puede resultar interesante la búsqueda de reglas de asociación en distintos niveles. Una forma habitual de especificar niveles de abstracción es el uso de jerarquías de clases, por lo que la búsqueda de reglas de asociación generalizadas es interesante cuando dichas jerarquías pueden ser aportadas por el usuario, en función de sus necesidades, o en el caso concreto de bases de datos orientadas a objetos.

- *Medidas para reglas de asociación.* Las medidas clásicas de soporte y confianza plantean algunos inconvenientes, relacionados principalmente con la presencia en la base de datos de items con muy alto soporte, y con la imposibilidad de detectar independencia estadística; se describen los inconvenientes del soporte y la confianza, proponiendo soluciones efectivas que, además de ser estadísticamente apropiadas, son más fáciles de comprender por el usuario.

- *Aplicaciones.* La extracción de reglas de asociación se ha utilizado para resolver diversos tipos de problemas y en distintos ámbitos, llegando a ser una de las técnicas más utilizadas. Uno de los motivos de esto es que los conceptos de item y transacción son conceptos abstractos, que se pueden hacer corresponder con distintos elementos de una base de datos en función de las necesidades del analista de los datos.

4.3 Ejemplos

Hablar de reglas de asociación implica entrar en detalle en cuanto a la generación de las mismas; la mayoría de los artículos publicados residentes en el Internet hacen referencia a la obtención de reglas de asociación a partir de un conjunto de datos o bien de una base de datos; para algunos autores consideran otras fuentes de datos para generar reglas de asociación, tal como son las series de tiempo.

Antes de entrar a detalle y hacer mención de la metodología usada en esta tesis para generar de reglas de asociación, pondremos un ejemplo de como obtener reglas a partir de un conjunto de datos.

Consideremos este grupo de elementos $I = \{A, B, C, D, E\}$ y un conjunto de transacciones $TID = \{100, 200, 300, 400\}$.

En la Tabla 4.1 los valores 100, 200, 300 y 400, son identificadores de cuatro transacciones: A=Azucar, B=Pan, C=Café, D=Leche y E=Pastel.

Podemos identificar reglas de asociación considerando el soporte y la confianza.

Sea:

minsupp=50% y *minconf*=60%. Al establecer estos parámetros se proponen dos pasos para obtener reglas de asociación.

El primer paso es contar las frecuencias de los k-items. En la Tabla 4.1, el elemento {A} ocurre en dos transacciones, cuyos identificadores son *TID*=100 y *TID*=300, presentando una frecuencia de 2, teniendo el soporte *supp(A)*=50% siendo este igual al mínimo soporte *minsupp*=50%.

TID	Items				
100	A		C	D	
200		B	C		E
300	A	B	C		E
400		B			E

Tabla 4.1: *Transacciones de una Base de datos*

El elemento {B} se hace presente en tres transacciones, teniendo los siguientes identificadores *TID*=200, *TID*=300 y *TID*=400, con un soporte *supp(B)*=75% siendo más grande al mínimo soporte.

Para el elemento {C}, de la misma manera que el elemento anterior tiene una ocurrencia en tres transacciones con los siguientes identificadores *TID*=100, *TID*=200 y *TID*=300, con un soporte *supp(C)*=75% siendo más grande al mínimo soporte.

Siguiendo con el elemento {D}, sólo presenta una ocurrencia con el siguiente identificador *TID*=100, con un soporte *supp(D)*=25%, menor al mínimo soporte.

Para el último elemento {E}, ocurre tres veces con los siguientes identificadores *TID*=100, *TID*=300 y *TID*=400 con un soporte *supp(E)*=75%, pasando el valor del mínimo soporte, resumido en la Tabla 4.2

Considerando 2-itemsets, en la Tabla 4.1, los elementos {A,B} ocurren en solo una transacción con el *TID*=300, teniendo una frecuencia igual a 1 y un soporte $supp(A \cup B) = 25\%$, siendo menor al mínimo soporte *minsupp*=50%.

El elemento {A,C} ocurre en dos transacciones con *TID*=100 y *TID*=300, con un soporte $supp(A \cup C) = 50\%$, siendo igual al mínimo soporte *minsupp*=50%.

Para el elemento {A,D} ocurre en una transacción con *TID*=100, con un soporte $supp(A \cup D) = 25\%$, siendo menor al mínimo soporte *minsupp*=50%.

De igual manera en el elemento {A,E} tiene sólo una ocurrencia en una transacción con *TID*=300, con un soporte $supp(A \cup E) = 25\%$, siendo menor al mínimo soporte *minsupp*=50%.

En este caso el elemento {B,C} ocurre en dos transacciones con *TID*=200 y *TID*=300, con un soporte $supp(B \cup C) = 50\%$, siendo igual al mínimo soporte *minsupp*=50% como se ve resumido en la Tabla 4.3.

De igual manera para 3-itemset y 4-itemset pueden ser obtenidos, como es listado en las Tablas 4.4 y 4.5.

Itemsets	Frecuencia > *minsupp*	
{A}	2	si
{B}	3	si
{C}	3	si
{D}	1	no
{E}	3	si

Tabla 4.2: *1-itemset en la Base de datos*

Itemsets	Frecuencia > *minsupp*	
{A,B}	1	no
{A,C}	2	si
{A,D}	1	no
{A,E}	1	no
{B,C}	2	si
{B,E}	3	si
{C,D}	1	no
{C,E}	2	si

Tabla 4.3: *2-itemset en la Base de datos*

Itemsets	Frecuencia > *minsupp*	
{A,B,C}	1	no
{A,B,E}	2	no
{A,C,D}	1	no
{A,C,E}	1	no
{B,C,E}	2	si

Tabla 4.4: *3-itemset en la Base de datos*

Itemsets	Frecuencia > *minsupp*	
{A,B,C,E}	1	no

Tabla 4.5: *4-itemset en la Base de datos*

El segundo paso es generar todas las reglas de asociación a partir de la frecuencia de los conjuntos de elementos(*itemsets*). Por consecuente al no existir alguna frecuencia en el itemset de la Tabla 4.5, el 4-itemset no contribuye con reglas de asociación válidas.

En la Tabla 4.4 sólo encontramos una frecuencia itemset {B,C,E}, con un soporte $supp(B \cup C \cup E) = 50\%$, siendo igual al mínimo soporte *minsupp*=50%, para esta frecuencia {B,C,E} donde $supp(B \cup C \cup E)/supp(B \cup C) = 2/2 = 100\%$ es mayor que la mínima confianza *minconf*=60%, $B \cup C \longrightarrow E$ se puede extraer como una regla válida.

De igual manera por el soporte $supp(B \cup C \cup E)/supp(B \cup E) = 2/3 = 66.7\%$, es mayor que la mínima confianza *minconf*=60%, $B \cup E \longrightarrow C$ se puede extraer como una regla válida.

A causa de que el soporte $supp(B \cup C \cup E)/supp(C \cup E) = 2/2 = 100\%$, es mayor que la mínima confianza *minconf*=60%, $C \cup E \longrightarrow B$ se puede extraer como una regla válida.

También al tener el soporte $supp(B \cup C \cup E)/supp(B) = 2/3 = 66.7\%$, es mayor que la mínima confianza *minconf*=60%, $B \longrightarrow C \cup E$ puede ser extraído como una regla válida.

No. de Regla	Regla	Confianza	Soporte	$> minconf$
Regla 1	$B \cup C \longrightarrow E$	100%	50%	si
Regla 2	$B \cup E \longrightarrow C$	66.7%	50%	si
Regla 3	$C \cup E \longrightarrow B$	100%	50%	si

Tabla 4.6: *Reglas de Asociación con 1-item consecuentes de 3-itemsets*

No. de Regla	Regla	Confianza	Soporte	$> minconf$
Regla 4	$B \longrightarrow C \cup E$	66.7%	50%	si
Regla 5	$C \longrightarrow B \cup E$	66.7%	50%	si
Regla 6	$E \longrightarrow B \cup C$	66.7%	50%	si

Tabla 4.7: *Reglas de Asociación con 2-item consecuentes de 3-itemsets*

Las reglas de asociación generadas de {B,C,E}se muestran en las Tablas 4.6 y Tabla 4.7.

También se pueden generar reglas de asociación desde la frecuencia de 2-itemsets de la Tabla 4.3, ilustrandose en las siguientes Tablas 4.8, 4.9, 4.10 y 4.11.

4.4 Reglas de Asociación basadas en Correlación

Un grupo de investigadores del Departamento de Ciencias de la Computación de la Universidad de Standard en [Brin et al,1997] retoman en caso del carrito de super-

No. de Regla	Regla	Confianza	Soporte	$> minconf$
Regla 7	$A \longrightarrow C$	100%	50%	si
Regla 8	$C \longrightarrow A$	66.7%	50%	si

Tabla 4.8: *Regla de Asociación para con* $\{A, C\}$

No. de Regla	Regla	Confianza	Soporte	$> minconf$
Regla 9	$B \longrightarrow C$	66.7%	50%	si
Regla 10	$C \longrightarrow B$	66.7%	50%	si

Tabla 4.9: *Regla de Asociación para con* $\{B, C\}$

No. de Regla	Regla	Confianza	Soporte	$> minconf$
Regla 11	$B \longrightarrow E$	100%	75%	si
Regla 12	$E \longrightarrow B$	100%	75%	si

Tabla 4.10: *Regla de Asociación para con* $\{B, E\}$

No. de Regla	Regla	Confianza	Soporte	$> minconf$
Regla 13	$C \longrightarrow E$	66.7%	50%	si
Regla 14	$E \longrightarrow C$	66.7%	50%	si

Tabla 4.11: *Regla de Asociación para con* $\{C, E\}$

mercado, con el propósito de generalizar reglas de asociación a reglas de correlación.

La idea básica que ellos establecen, consiste en que las reglas de asociación no capturan dependencias interesantes entre los elementos, como el caso de "que alguien que compra café usualmente no compra té"; la clave está en que al comprar café y comprar té no hay una asociación pero si una correlación, aunque en este caso "correlación" es impropio, y mejor sería usar "dependencia" ya que la correlación es usado para hechos históricos.

Establecen que dado dos eventos A, B son independientes:
if $prob(A \cap B) = prob(A)prob(B)$, cuyo caso contrario existe una dependencia. El procedimiento estadístico para probar la independencia es mediante la prueba de *Chi Cuadrada*.

El alcance que presentan este tipo de reglas es que consideran a los *ítems* como un conjuntos de variables, las transacciones son considerados como observaciones y cada *itemset* usa la prueba de *Chi Cuadrada* para inferir a partir del conjunto de transacciones, si estos *ítems* son dependientes.

En [Batyrshin et al,2004] definen una regla basada en correlación bajo la siguiente estructura:

If Cond then A está asociado con B, (W)

Donde:

Cond es una restricción en series de tiempo, y la *asociación* es calculado como un coeficiente de correlación entre dos restricciones de las series de tiempo A y B, el cual describen algunos parámetros del sistema analizado. W es una significancia de la regla dado los valores del coeficiente de correlación y de t-test. A continuación se muestra un ejemplo de este tipo de asociación:

If producción de aceite (PozolT101)= es Alto then producción de aceite(PozolT25) es altamente asociado con la producción de gas(PozoT03), (W).

4.5 Otros tipos de Reglas de Asociación

Encontramos otro tipo de reglas de asociación, tal es el caso de las *Reglas de Asociación Difusas*, éste tipo de reglas utilizan conceptos sobre conjuntos difusos, de tal forma que atributos cuantitativos pueden ser manejados.

Considere el siguiente conjunto $T = \{t_1, t_2, ..., t_n\}$, que representa una base de datos y t_i representa las i-ésima tupla en T, usaremos $I = \{i_1, i_2, ..., i_m\}$ para representar todos los atributos que aparecen en T e i_j representa la j-ésimo atributo. I contiene todos los conjuntos de elementos y llamaremos a I un *itemset*.

Retirado	**No de Hijos**	**Salario**
si	2	0
no	3	15000
no	0	10000
no	1	20000
si	2	0

Tabla 4.12: *Una simple Base de Datos*

En la Tabla 4.12 se muestra una simple base de datos con atributos cuantitativos.

Tenemos que $T = \{t_1, t_2, t_3, t_4, t_5\}$ e $I = \{Retirado, No\ de\ Hijos, Salario\}$, podemos conocer el valor del atributo i_k del registro j-ésimo simplemente por $t_j[i_k]$. Por ejemplo, si queremos conocer el valor del salario del cuarto registro, solamente usamos $t_4[Salario]$, devolviendo el valor de 20000.

Para cada atributo i_k estará asociado con diversos conjuntos difusos. Usamos $F_{i_k} = \{f_{i_k}^1, f_{i_k}^2, \ldots, f_{i_k}^l\}$ para regresentar el conjunto difuso asociado con i_k y $f_{i_k}^j$ representa el j-ésimo conjunto difuso en F_{i_k}. Por ejemplo, si el atributo Salario tiene tres conjuntos difusos: alto, medio y bajo , tendremos $F_{Salario} = \{alto, medio, bajo\}$. Los conjuntos difusos y su correspondiente función de membresía son provistos por expertos en el dominio.

Dada una base de datos T con atributos I y estos conjuntos difusos asociados con atributos en I, queremos encontrar algo interesante y potencialmente útil en una forma metodológica. Se propone la siguiente regla de asociación difusa:

If X está A then Y está en B:

En base a la regla descrita ateriormente, considere a $X = \{x_1, x_2, \ldots, x_p\}$ y $Y = \{y_1, y_2, \ldots, y_q\}$ como *itemsets*. X y Y son subconjuntos de I y disjuntos, es decir, que no comparten atributos comunes. $A = \{f_{x_1}, f_{x_2}, \ldots, f_{x_p}\}$ y $B = \{f_{y_1}, f_{y_2}, \ldots, f_{y_q}\}$ contienen los conjuntos difusos asociados con sus correspondientes atributos en X y Y. Por ejemplo, un atributo x_k en X tendrá un conjunto difuso f_{x_k} en A tal que $f_{x_k} \in F_{x_k}$ sea satisfecho.

La primer parte de la regla "X está en A" es llamado *antecente* y "Y está en B" es llamado *consecuente* de la regla. La semántica de la regla es cuando "X está en A" es satisfecho, podemos implicar que "Y está en B" y también es satisfecho. Por

lo tanto, *satisfecho* significa que hay una cantidad suficiente de registros los cuales contribuyen sus votos al atributo-conjunto de pares difusos y la suma de esos votos es mayor que el umbral introducido por el usuario. Si una regla es interesante, ésta tendrá suficiente significado y un alto factor de certidumbre. Usaremos el significado y el valor de certidumbre para determinar la veracidad de los *itemsets* y las reglas [Chan & Au,1997].

4.6 Metodología de generación de reglas de asociación basadas en correlación usando el paradigma de agentes

El presente apartado tiene el propósito de dar a conocer los pasos de la metodología propuesta para generar reglas de asociación basados en análisis de correlación mediante el uso del paradigma de agentes.

La metodología consiste de los siguientes pasos, tal como lo muestra la Figura 4.1:

1. *Selección*: Esta primera parte, es aquella donde se hace la selección de las variables a analizar.
2. *Preprocesamiento*: En esta fase se establecen valores lingüísticos, los cuales serán la base para la condición de las reglas de asociación; se definen los intervalos que cumplen con la condición.
3. *Procesamiento*: Es la etapa donde se lleva a cabo el análisis de correlación mediante cálculos estadísticos implicando el soporte de las reglas para finalmente obtener las reglas de asociación.

A continuación se describirá en detalle cada una de las partes mostradas de la metodología de la Figura 4.1.

Como primer punto tenemos los *datos de entrada*, está es la fase inicial de la metodología, aquí se lleva a cabo la lectura de los datos que se encuentran físicamente distribuidos; entiéndase como datos de entrada las *series de tiempo* a analizar, aunado a ello es importante considerar la *variable* de entrada que de igual manera se va a analizar, para nuestro caso de estudio descrito en el capítulo 6, está variable puede ser: gas, agua o aceite.

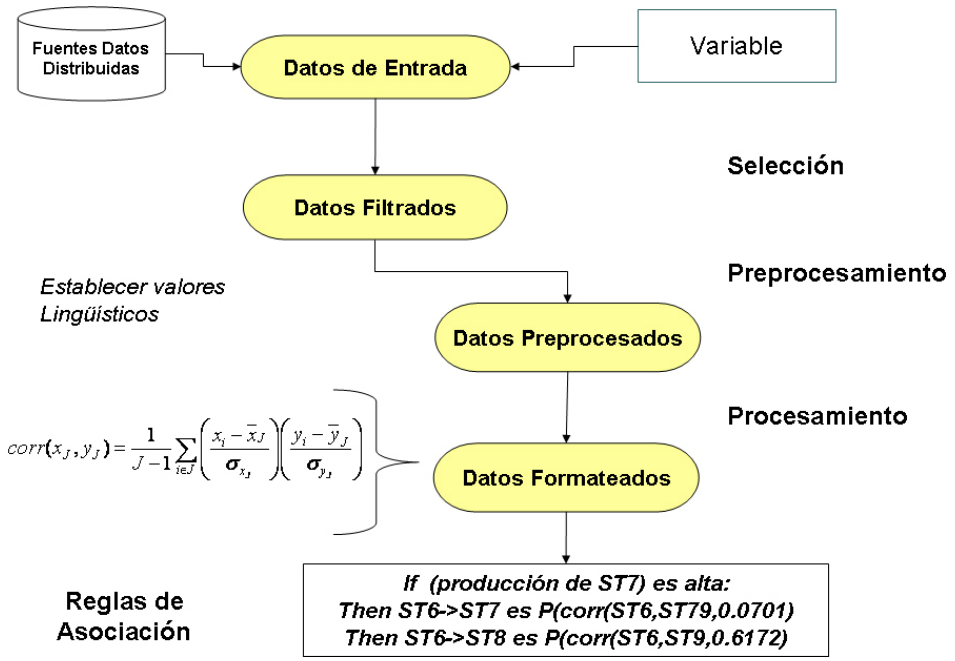


Figura 4.1: *Pasos generales de la metodología para generar reglas de asociación*

Llegamos a la etapa denominado *Selección* descrito líneas atrás, cuya importancia radica en la selección de los datos que deseamos analizar en base a la variable elegida con anterioridad y es por ello que lo denominamos como *datos filtrados* ya que de todas las series de tiempo cargados en el sistema, se hace un selección o una muestra de todos los datos.

El *Preprocesamiento* es la siguiente etapa de la metodología propuesta, la cual tiene el objetivo de establecer la condición de las reglas, ésta condición es determinado por los *valores lingüísticos* definidos al seleccionar una serie de tiempo base para el análisis de correlación; poniendo un ejemplo de valores lingüísticos consideramos: una producción baja, media o alta de cierta variable definida en los datos de entrada. Otro punto importante en esta fase de la metodología, es el hecho de extraer fragmentos de las series de tiempo que cumplan con la condición definida con anterioridad, es por ello que le denominamos *datos preprocesados*, ya que las series de tiempo han pasado por una etapa de selección, se ha establecido la condición de las reglas y extraído fragmentos de dichas series de tiempo en base a una condición.

En el *Procesamiento* se establece el soporte de las reglas de asociación a través del análisis de correlación, éste análisis se hace por medio de *cálculos estadísticos* tomando en cuenta las series de tiempo que intervienen en ello como una matriz de

datos, la cual es pasada por varias operaciones hasta obtener la *matriz de coeficientes de correlación*. Una vez obtenida la matriz de coeficientes de correlación, es decir, datos formateados, se interpreta la matriz resultante generando finalmente las *reglas de asociación*.

El uso del paradigma de agentes dentro de la metodología propuesta, se debe a que las características que presenta ésta tecnología y la arquitectura basada en servicios, hará que cada de los pasos descritos líneas atrás vendrán a formar parte de los servicios que cada agente de nuestro SMA proporcionará. Los agentes identificados y los servicios que cada uno de ellos ofrece, están descritos en el siguiente capítulo.

Capítulo 5

Análisis y diseño del SMA para el análisis de datos de series de tiempo y extracción de reglas de asociación

En el presente Capítulo se hace un tratado del análisis y diseño del sistema multi-agente para llevar a cabo el análisis de series de tiempo y la extracción de reglas de asociación haciendo uso del paradigma de agentes, considerando la metodología descrita en el capítulo anterior.

5.1 Análisis y diseño del SMA

En este apartado se presenta el Análisis y Diseño del SMA del prototipo que se ha implementado, para lo cual el SMA tiene que llevar considerar los siguientes aspectos:

1. El sistema tiene que hacer el análisis de los datos (es decir series de tiempo), tomando en cuenta que los datos pueden encontrarse en fuentes distribuidas.

2. Basándose en la metodología propuesta en el capítulo anterior nuestro sistema debe permitir la selección de series de tiempo que sean de interés para el usuario, estableciendo parámetros que sean de interés para el análisis, así también como delimitar o seleccionar intervalos de las series de tiempo que han sido elegidas por parte del usuario.

3. Un aspecto importante es obtener los cálculos estadísticos a partir de las series de tiempo seleccionadas para poder obtener una matriz de coeficientes de corre-

lación y en base a ello generar reglas de asociación que es lo que nuestro SMA desea obtener.

Una vez definido lo que nuestro sistema debe de requerir, a continuación pasamos a la parte del Análisis y Diseño del SMA.

5.1.1 Análisis

Como parte primordial, se describen las especificaciones del sistema mediante los diagramas de casos de uso, el modelo de agentes, de servicios e interacciones.

5.1.1.1 Casos de Uso

En la Figura 5.1, se muestra el diagrama de caso de usos del sistema multiagente.

Como se puede observar, se describe de manera general las actividades importantes llevadas acabo por el SMA, tales como iniciar la plataforma de desarrollo y el SMA, hasta visualizar los resultados obtenidos; para nuestro caso, es la visualización de reglas de asociación generados a partir del análisis de series de tiempo.

5.1.1.2 Modelo de Agentes (MA)

Es en esta sección es donde describimos los agentes que van a intervenir en nuestro SMA, los cuales identificamos los siguientes: *Agente Datos* (ver Tabla 5.1), *Agente Correlación* (ver Tabla 5.2), *Agente GraficaST* (ver Tabla 5.3), *Agente GeneradorRA* (ver Tabla 5.4).

5.1.1.3 Modelo de Servicios (MS)

Este modelo permite conocer los servicios que serán implementados por el SMA que se propone en la presente tesis, entre los cuales tenemos los siguientes: *LeerDatos* (ver Tabla 5.5), *SeleccionarDatos* (ver Tabla 5.6), *VisualizarDatos* (ver Tabla 5.7), *DefinirCondiciónReglas* (ver Tabla 5.8), *GenerarMatrizdeCorrelación* (ver Tabla 5.15), *GenerarReglasAsociación* (ver Tabla 5.10).

5.1.1.4 Modelo de Interacciones (MI)

Este modelo se construye a partir del análisis del sistema considerando el diagrama de caso de uso y el modelo de servicios para lo cual se tiene que considerar que este modelo está basado en la comunicación existente entre los agentes del sistema.

Nombre	**Agente Datos**
Servicios	LeerDatos
Funcionalidad General	Este Agente es el que permite la lectura de datos (en este caso series de tiempo), las cuales son seleccionadas por el usuario, es el agente que interactúa con el usuario a través de una interfaz. También proporciona los datos leídos a otros agentes para que estos puedan trabajar con ellos.
Ontologías	Ninguna

Tabla 5.1: *Descripción Agente Datos*

Nombre	**Agente Correlación**
Servicios	Hacer el análisis de correlación, generar matriz de correlación, cálculos estadísticos
Funcionalidad General	Este Agente realiza diversas actividades las cuales son de gran importancia, ya que el hace una análisis de correlación de las series de tiempo seleccionadas por el usuario, generando una matriz de correlación en base a cálculos estadísticos que este agente realiza.
Ontologías	Ninguna

Tabla 5.2: *Descripción Agente Correlación*

Nombre	**Agente GraficaST**
Servicios	Graficar Series de Tiempo
Conocimiento inicial	Series de tiempo seleccionadas por el usuario y fragmentos de series de tiempo.
Funcionalidad General	La funcionalidad principal de éste Agente, es representar de manera gráfica las series de tiempo y fracciones de series de tiempo seleccionados por parte del usuario.
Ontologías	Ninguna

Tabla 5.3: *Descripción Agente GraficaST*

Nombre	**Agente GeneradorRA**
Servicios	Generar Reglas de Asociación
Conocimiento inicial	Matriz de correlación
Funcionalidad General	Como su nombre lo indica, es el Agente encargado de generar las reglas de Asociación correspondientes, basándose en el análisis de las series de tiempo.
Ontologías	Ninguna

Tabla 5.4: *Descripción Agente GeneradorRA*

Nombre	**LeerDatos**
Clase de agente	AgenteDatos
Conocimiento de entrada	LecturaDatos(series de tiempo)
Conocimiento de salida	EnviarDatos(series de tiempo)
Descripción de la funcionalidad e implicaciones internas en el agente y externas en el SMA	Este servicio permite a los agentes obtener los datos iniciales, es decir, las series de tiempo con las que va a estar interactuando el resto de los agentes existentes en el sistema.
Ontologías	Ninguna

Tabla 5.5: *Descripción del servicio LeerDatos*

Nombre	**SeleccionarDatos**
Clase de agente	AgenteDatos
Conocimiento de entrada	LecturaDatos(series de tiempo)
Conocimiento de salida	EnviarDatos(series de tiempo)
Descripción de la funcionalidad e implicaciones internas en el agente y externas en el SMA	Al igual que el servicio descrito anteriormente, se lleva a cabo la lectura de los datos, considerando el hecho de permitirle al usuario el poder elegir los datos que únicamente considere útiles para el análisis.
Ontologías	Ninguna

Tabla 5.6: *Descripción del servicio SeleccionarDatos*

Nombre	**VisualizarDatos**
Clase de agente	Agente GraficaST
Conocimiento de entrada	LecturaDatos(series de tiempo)
Conocimiento de salida	EnviarDatos(series de tiempo)
Descripción de la funcionalidad e implicaciones internas en el agente y externas en el SMA	Este servicio permite representar de manera gráfica a través de una interfaz las series de tiempo que fueron capturados por el AgenteDatos.
Ontologías	Ninguna

Tabla 5.7: *Descripción del servicio VisualizarDatos*

Nombre	**DefinirCondiciónReglas**
Clase de agente	Agente GraficaST
Conocimiento de entrada	valoresLinguisticos(valor), definirIntervalo(rango)
Conocimiento de salida	VisualizarValoresLinguisticos(valor, rango)
Descripción de la funcionalidad e implicaciones internas en el agente y externas en el SMA	El propósito de este servicio es definir las condiciones de las reglas de asociación a generar, considerando los valores lingüísticos establecidos por el usuario, y representarlos en la gráfica de las series de tiempo.
Ontologías	Ninguna

Tabla 5.8: *Descripción del servicio DefinirCondiciónReglas*

Nombre	**GenerarMatrizCorrelación**
Clase de agente	Agente GeneradorRA
Conocimiento de entrada	LecturaDatos(Series de tiempo que cumplen con la condición)
Conocimiento de salida	generarMatriz(matrizCorrelación)
Descripción de la funcionalidad e implicaciones internas en el agente y externas en el SMA	En este servicio se produce una matriz de correlación, considerando las series de tiempo que cumplen con la condición establecida por los valores lingüísticos, para lo cual se llevan a cabo cálculos estadísticos para la generación de la matriz de correlación.
Ontologías	Ninguna

Tabla 5.9: *Descripción del servicio GenerarMatrizCorrelación*

Nombre	**GenerarReglasAsociación**
Clase de agente	Agente GeneradorRA
Conocimiento de entrada	LecturaDatos(matrizCorrelación)
Conocimiento de salida	generarReglasAsociacion(ReglasAsociación)
Descripción de la funcionalidad e implicaciones internas en el agente y externas en el SMA	La razón primordial de este servicio, es la generación de las reglas de asociación, basados en un análisis de correlación hecho previamente por los servicios anteriores.
Ontologías	Ninguna

Tabla 5.10: *Descripción del servicio GenerarReglasAsociación*

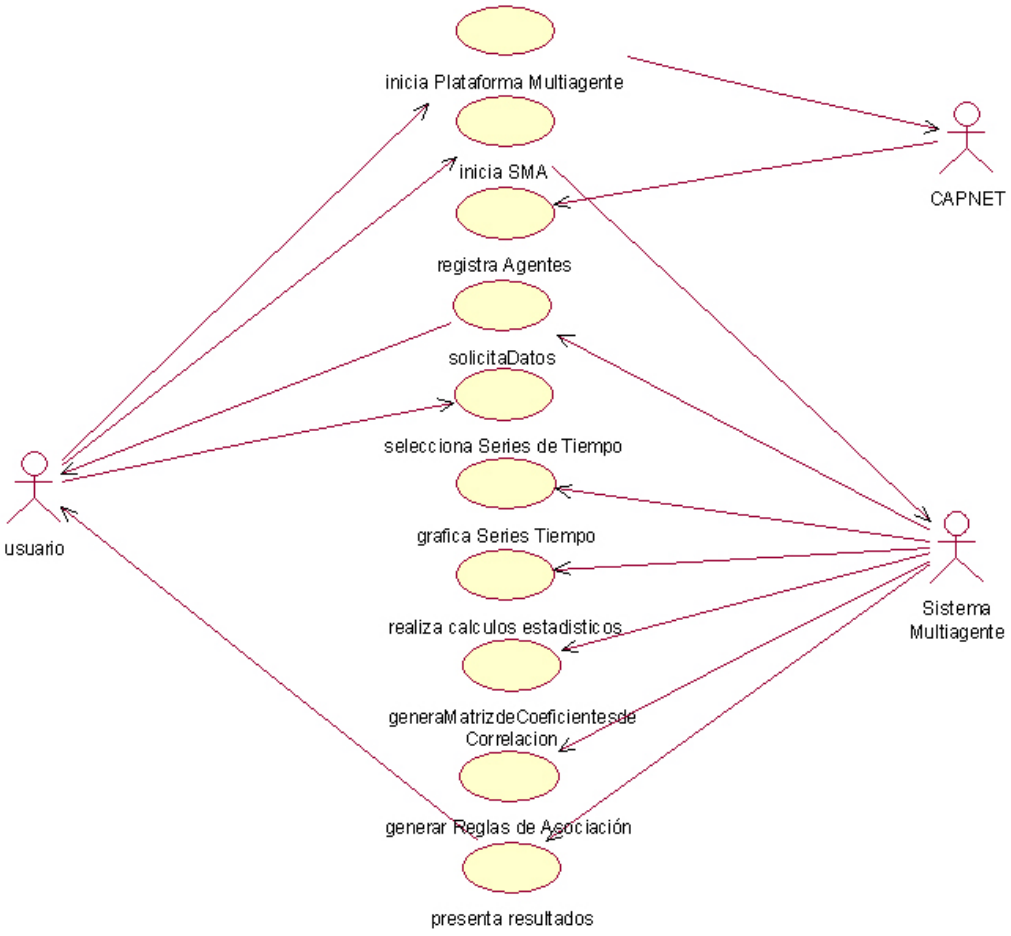


Figura 5.1: *Diagrama de Casos de Uso del SMA*

Las interacciones identificadas en este modelo son las siguientes: *SolicitarRegistro* (ver Tabla 5.11), *SolicitarDatos* (ver Tabla 5.12), *SeleccionarDatos* (ver Tabla 5.13), *GraficarDatos* (ver Tabla 5.14), *GenerarMatrizCorrelación* (ver Tabla 5.15) y *GenerarReglasAsociación* (ver Tabla 5.16).

Nombre	**SolicitarRegistro**
Agentes	AgenteDatos, AgenteCorrelacion, AgenteGraficaST, AgenteGeneradorRA
Mensajes	Request, Inform
Secuencia	Fipa-Request Protocol
Descripción	Por medio de esta interacción, los agentes del SMA, solicitan el registro de los mismos en la plataforma de desarrollo de agentes, en nuestro caso de CAPNET, para mantener la existencia de los agentes y del registro de los servicios que ofrece cada agente.

Tabla 5.11: *Descripción de la interacción SolicitarRegistro*

Nombre	**SolicitarDatos**
Agentes	AgenteDatos
Mensajes	Request, Inform
Secuencia	
Descripción	En esta interacción, el AgenteDatos, establece una comunicación con el usuario, solicitando los datos correspondientes a analizar (Series de Tiempo), en otras palabras es la interacción donde se "cargan" los datos a analizar.

Tabla 5.12: *Descripción de la interacción SolicitarDatos*

Nombre	**SeleccionarDatos**
Agentes	Usuario, AgenteDatos, AgenteGraficaST
Mensajes	Request, Inform
Secuencia	
Descripción	El usuario desempeña un papel importante en esta interacción debido a que el AgenteGraficaST requiere que el usuario establezca la condición de las reglas de asociación a generar seleccionando los datos (Series de tiempo) que intervendrán en la condición.

Tabla 5.13: *Descripción de la interacción SeleccionarDatos*

Nombre	**GraficarDatos**
Agentes	AgenteDatos, AgenteGraficaST
Mensajes	Inform
Secuencia	
Descripción	Esta interacción permite al AgenteGraficaST representar de manera gráfica los datos (Series de Tiempo) que el usuario previamente ha seleccionado, de igual manera, se va reflejando los cambios que se van suscitando a través de todo el procesamiento de los datos.

Tabla 5.14: *Descripción de la interacción GraficarDatos*

Nombre	**GenerarMatrizCorrelación**
Agentes	AgenteCorrelacion, AgenteGraficaST
Mensajes	Request, Inform
Secuencia	Fipa-Request Protocol
Descripción	En esta interacción, se solicita generar una matriz de correlación en donde el AgentGraficaST tiene la condición definida por el usuario y estos datos son enviados al AgenteCorrelacion, el cual se realiza mediante el cálculo de datos estadísticos para poder generar la matriz.

Tabla 5.15: *Descripción de la interacción GenerarMatrizCorrelación*

Nombre	**GenerarReglasAsociación**
Agentes	AgenteCorrelacion, AgenteGeneradorRA
Mensajes	Request, Inform
Secuencia	Fipa-Request Protocol
Descripción	La interacción existe entre estos agentes, donde el AgenteCorrelacion envía la matriz generada al AgenteGeneradorRA y de la misma manera le solicita la generación de las reglas de asociación que es de interés en nuestro SMA.

Tabla 5.16: *Descripción de la interacción GenerarReglasAsociación*

5.1.2 Diseño

5.1.2.1 Modelo de Agentes (MA)

En este modelo los agentes identificados y las plantillas usadas son representados por medio de un diagrama de clases en UML.

El diagrama de clases mostrado en la Figura 5.2, incluye la clase *AgentBasic*, que es una clase que tomamos como platilla de los agentes que conforman nuestro SMA. La clase *ZedGraph* es otra plantilla que se utiliza el *Agente GraficaST*, el cual permite la visualización de las series de tiempo (datos) que son manipulados por todo el proceso descrito en la metodología del capítulo anterior. La clase *BasicStatistics* es una clase que de igual forma es utilizado como una platilla, usado por el *Agente Datos* y el *Agente Correlación*. La clase *Correlation* es una plantilla usado por el *Agente Correlación* y el *Agente GeneradorRA*, el cual permite llevar a cabo el análisis de correlación entre las series de tiempo seleccionadas en nuestro SMA.

5.1.2.2 Modelo de Servicios (MS)

En este modelo, como se mencionó en la sección del Análisis del SMA, permite considerar el detalle de las actividades que son ejecutadas cuando un servicio de un agente es ejecutado; mediante el uso de los diagramas de actividades conoceremos en detalle cada servicio.

Los servicios identificados y su respectivo diagrama de actividad son los siguientes:

- *LeerDatos*: Para que pueda llevarse a cabo este servicio requiere ser invocado, motivo por el cual espera la solicitud de lectura de datos, en caso de aceptarse la solicitud la lectura de datos es llevada a cabo directamente en las bases de datos correspondientes ubicadas en los diferentes sitios, en caso contrario se presenta un mensaje de error (Ver Figura 5.3).

- *SeleccionarDatos*: En este servicio el usuario interviene en la selección de las series de tiempo a analizar, el Agente LeerDatos, se encarga de enviar esos datos al Agente asistente personal de usuario para ser mostrados en una interfaz gráfica de usuario (GUI) y así seleccionar los datos correspondientes; este servicio se realiza si los datos seleccionados son entregados de forma correcta, en caso contrario se presenta un mensaje de error (Ver Figura 5.4).

- *VisualizarDatos*: Su propósito es mostrar de manera gráfica los datos que son recibidos por este servicio, al ser invocado estos datos son validados y se procede a representarlos de manera gráfica, en caso contrario se presenta una excepción (Ver Figura 5.5).

- *DefinirCondiciónReglas*: El usuario juega un papel importante en este servicio, ya que él tiene la responsabilidad de definir la condición de las reglas, estos parámetros son enviados al Agente GraficaST a través de la GUI presentado por el Agente asistente personal de usuario; los datos son validados y en caso de haberlos recibido correctamente, se procede a representar estos parámetros en la interfaz, de lo contrario se manda un mensaje de error (Ver Figura 5.6).

- *GenerarMatrizCorrelacion*: En este servicio una vez que ha sido invocado, se leen los parámetros definidos por el servicio descrito con anterioridad, en caso de ser válidos se llevan a cabo los cálculos estadísticos correspondientes para que posteriormente se genere una matriz de correlación descrito en el próximo capítulo (Ver Figura 5.7).

- *GenerarReglasAsociacion*: Para llevarse a cabo el presente servicio, es necesario leer la matriz de coeficientes de correlación generado por el Agente Correlación, si no se presenta alguna anomalía en la lectura de dicha matriz, se procede a la generación de reglas de asociación basados en correlación, en caso contrario se presenta un mensaje de error (Ver Figura 5.8).

5.1.2.3 Modelo de Interacciones (MI)

En este apartado se describen las interacciones entre los diferentes agentes existentes en el sistema, las cuales están descritas en la sección de Análisis del presente capítulo; por medio de un diagrama de interacciones, se muestra de forma gráfica los mensajes que intervienen entre los agentes (ver Figura 5.9) y a través de un diagrama de colaboración, es posible ver la forma en que cada una de las clases de nuestro sistema, se comunican entre ellas (ver Figura 5.10).

5.1.2.4 Modelo SMA

El propósito de este modelo es describir la arquitectura del SMA que ha sido usado para la implementación del sistema abordado en la presente tesis; ésta descripción tiene como fundamento el describir la Plataforma de Agentes utilizado, presentando la distribución física de los componentes del sistema mediante un diagrama de despliegue.

Plataforma de Agentes: CAPNET

Una plataforma de agentes, es una arquitectura de software que controla y administra una comunidad de agentes, permitiendo a los agentes la supervivencia y movilidad en un ambiente distribuido y heterogéneo. En el capítulo 2, se hace mención de las plataformas de desarrollo para sistemas multi-agente, entre ellos aparece CAPNET.

CAPNET es la plataforma de agentes elegida para el desarrollo de nuestro SMA, ya que nos brinda un entorno de desarrollo (ver Figura 2.2), el cual permite obtener productos derivados basados en una arquitectura orientada en servicios.

Diagrama de Despliegue del sistema

La presente tesis forma parte de un modulo de un sistema que el IMP está desarrollando denominado Smart-Agua [Leonid et al,2006], la cual comprende de distintos módulos; la infraestructura del sistema se basa en la plataforma para el despliegue de Sistemas Multi-agentes denominada CAPNET, del cual se habló en la capítulo 2.

La arquitectura de Smart-Agua, como se muestra en la Figura 5.11, incluye el sistema experto que conforma la capa de aplicación de Smart-Agua, esto incluye la Máquina CAPNET de Inferencia Difusa, las bases de conocimiento y los módulos de minería de datos; en su conjunto representan la parte principal del sistema. La máquina de inferencia difusa se invoca por la aplicación principal cada vez que se requiere aplicar algún mecanismo de razonamiento. La herramienta de minería de datos forma parte del Toolbox de Minería de Series de Tiempo Perceptual - Percept-Miner. El módulo de cálculo comparte datos con la máquina de inferencia y se encarga de sugerir valores en algunos puntos durante la localización de soluciones. Se esta previendo la integración del sistema experto con simuladores como Coning-Frac (flujo multifásico en pozos de yacimientos fracturados).

La alimentación de datos en el estado actual de desarrollo del sistema se realiza mediante la importación de tablas contenidas en formatos externos como hojas de cálculo Microsoft Excel y archivos de texto; enfocándonos en el módulo de Minería de Datos, la cual es parte esencial de esta tesis y en la que nos hemos enfocado, es aquí donde los agentes que se identificaron en las secciones anteriores, tienen razón de ser de su existencia y son los siguientes:

1. Agente Datos.
2. Agente GraficaST.
3. Agente Correlación.
4. Agente GeneradorRA.

Dichos agentes fueron construidos con la plataforma CAPNET; en la Figura 5.12 se muestra un diagrama básico de la arquitectura de la solución propuesta, cuya descripción de cada agente está reflejado en secciones anteriores, éstos agentes son registrados en el AMS y los servicios que cada uno de ellos brinda, son registrados en el DF para ser invocados cuando sean necesarios.

Figura 5.2: *Diagrama de clases del SMA*

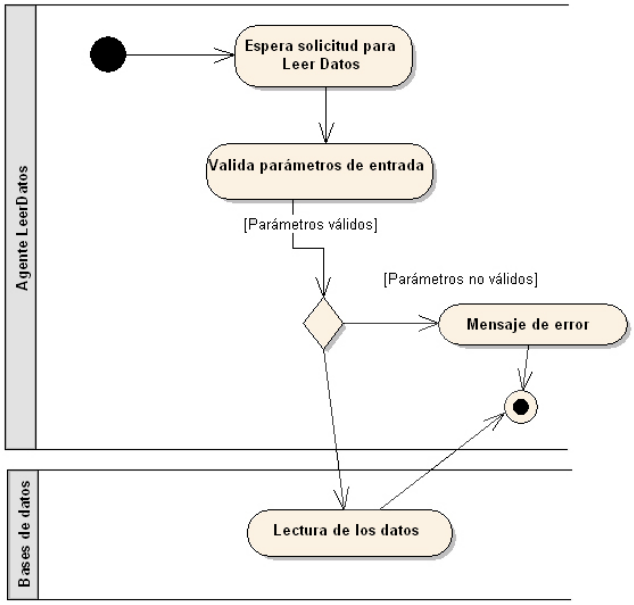


Figura 5.3: *Diagrama de Actividades del Servicio LeerDatos*

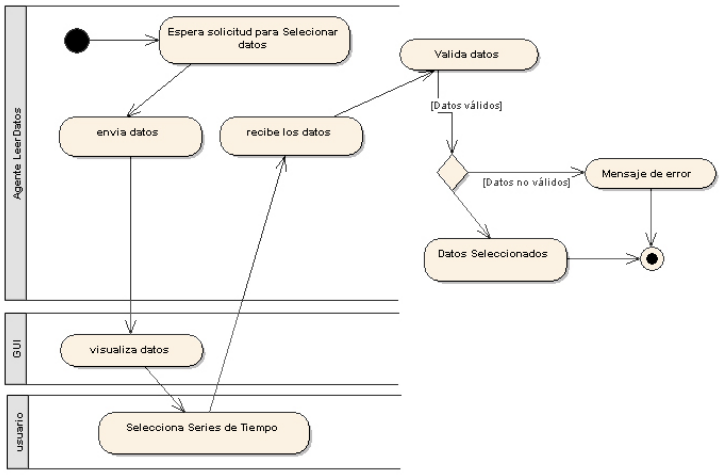


Figura 5.4: *Diagrama de Actividades del Servicio SeleccionarDatos*

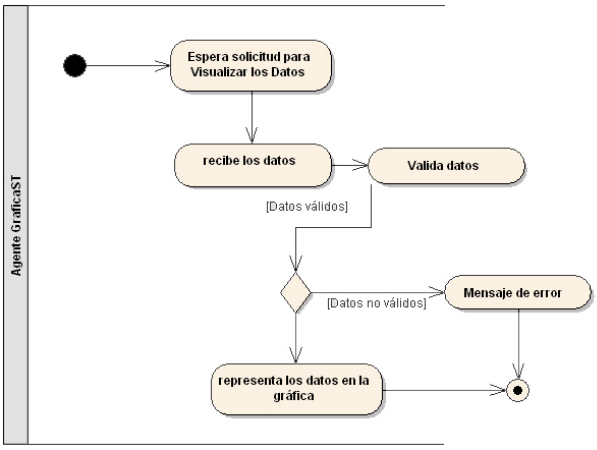


Figura 5.5: *Diagrama de Actividades del Servicio VisualizarDatos*

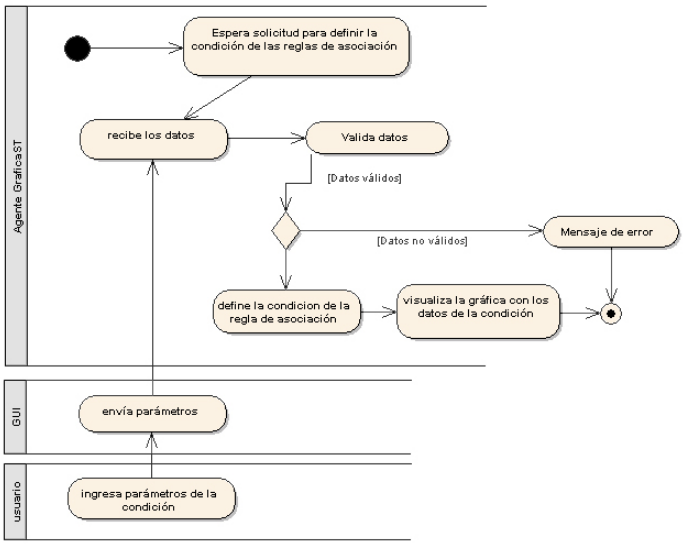


Figura 5.6: *Diagrama de Actividades del Servicio DefinirCondiciónReglas*

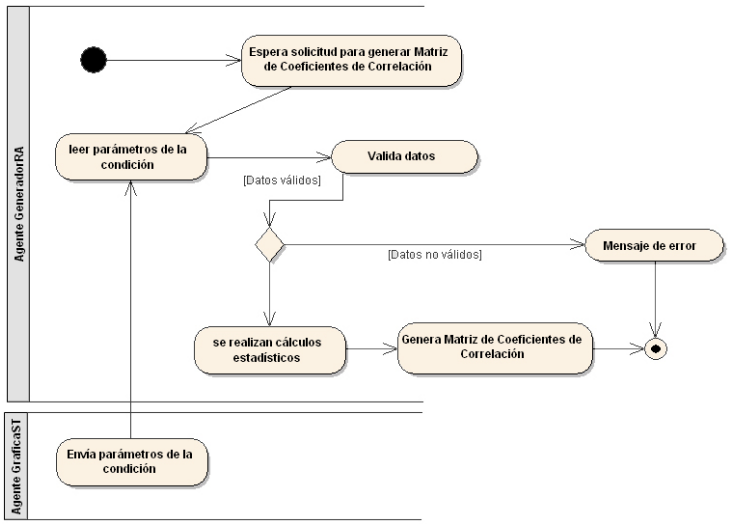


Figura 5.7: *Diagrama de Actividades del Servicio GenerarMatrizCorrelación*

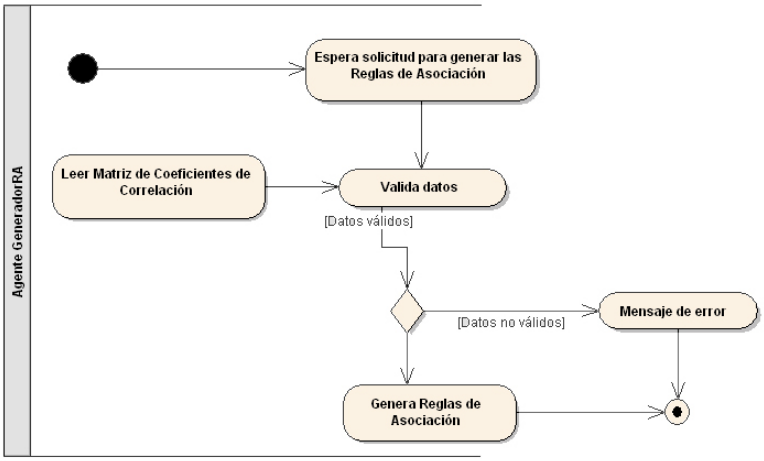


Figura 5.8: *Diagrama de Actividades del Servicio GenerarReglasAsociación*

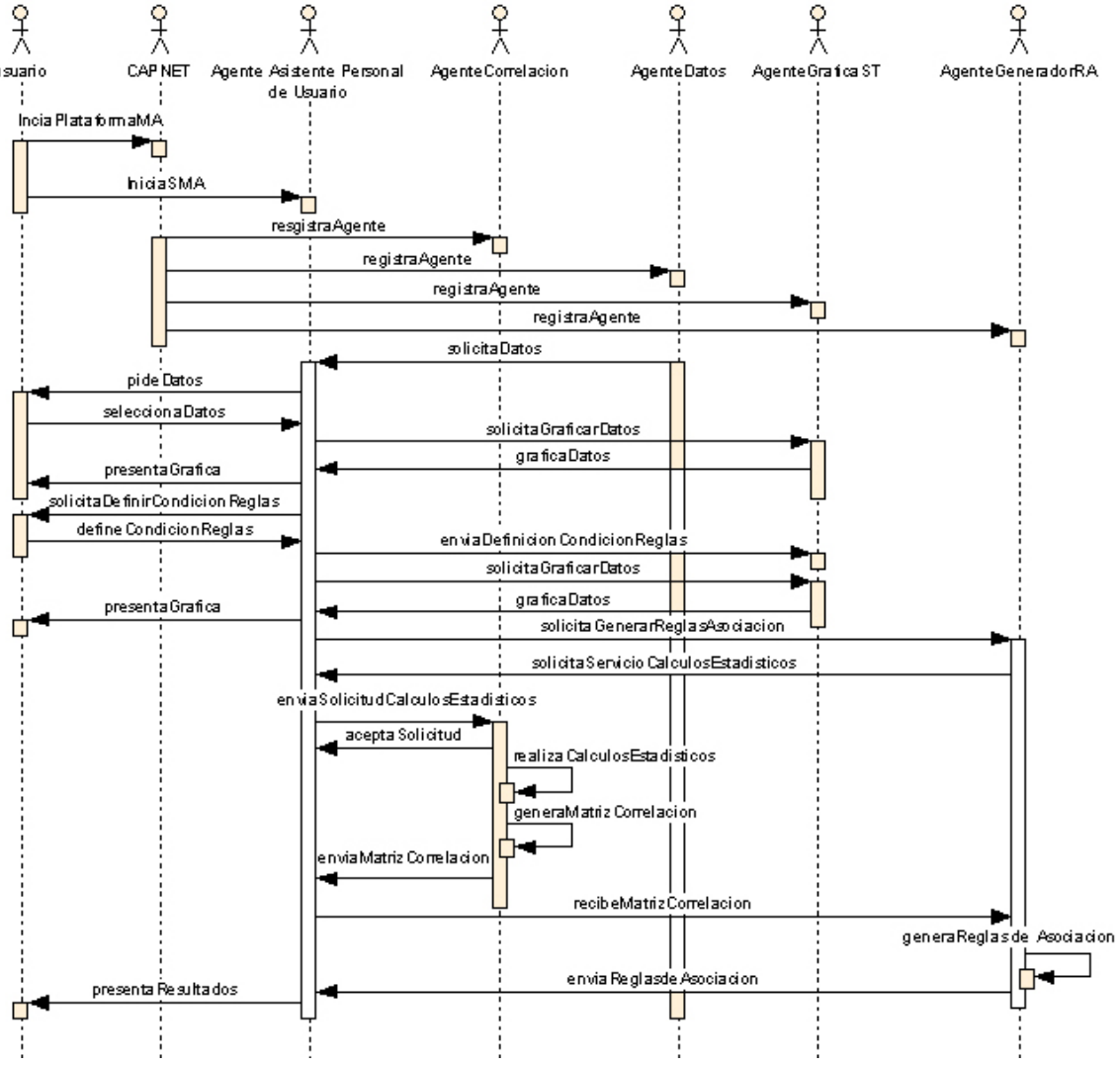


Figura 5.9: *Diagrama de Interacciones del SMA*

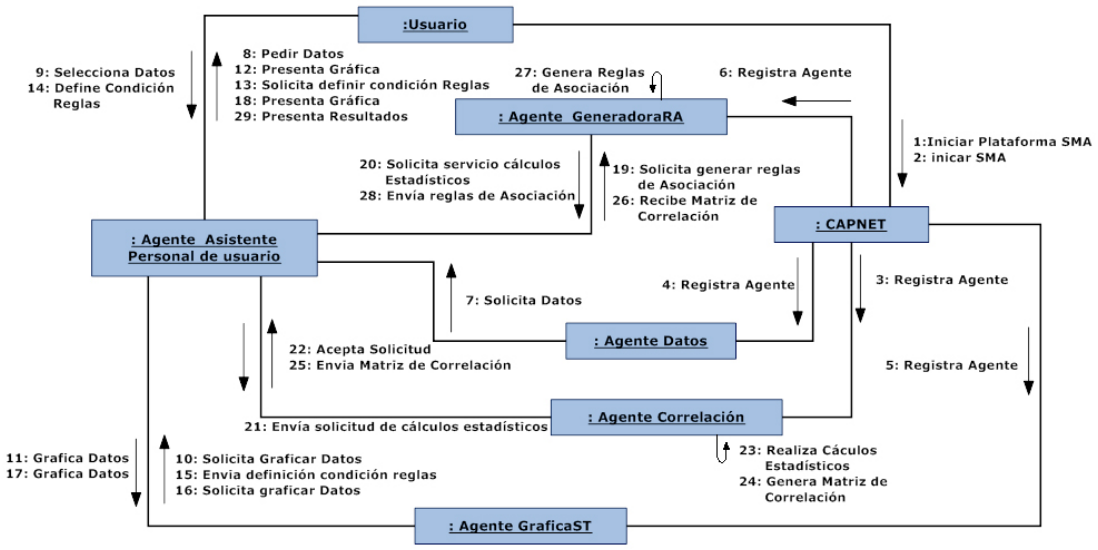


Figura 5.10: *Diagrama de Colaboración del SMA*

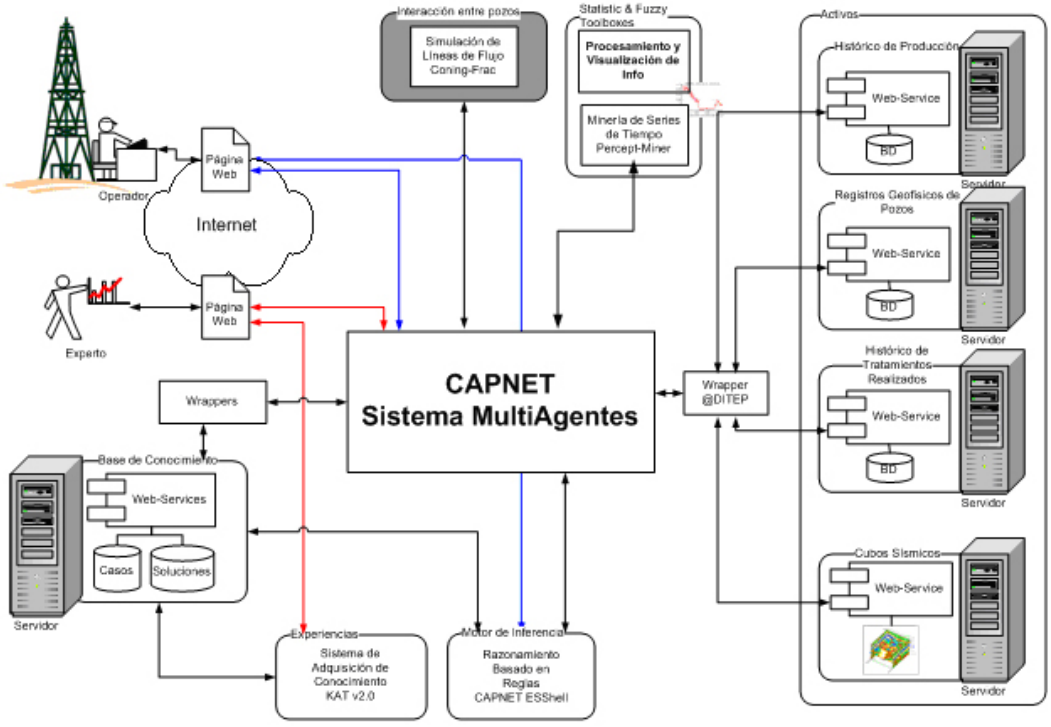


Figura 5.11: *Arquitectura de Smart-Agua*

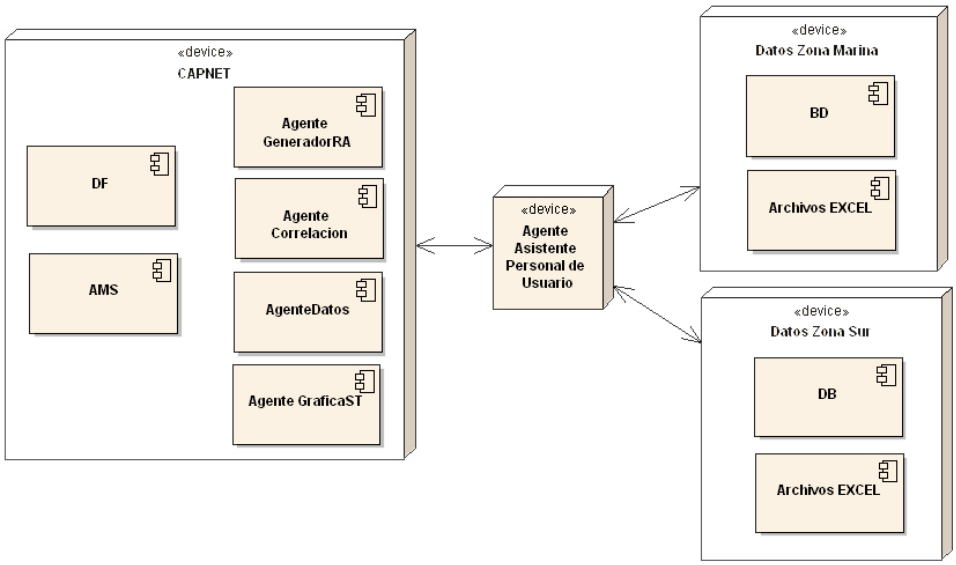


Figura 5.12: *Arquitectura del Sistema Multi-Agente Propuesto*

Capítulo 6

Resultados Experimentales

En el presente Capítulo se dan a conocer los resultados obtenidos al implementar un SMA que hace uso de la metodología propuesta en el capítulo anterior para extraer de reglas de asociación; se hace una descripción del caso de estudio a implementar mostrando los resultados del mismo como se definió anteriormente.

6.1 Introducción

El propósito de este apartado y del presente trabajo de tesis es aplicar la metodología propuesta, la cual es implementada sobre el SMA desarrollado para un caso en particular; cada una de las fases de la metodología (ver Figura 4.1), fueron implementados en cada uno de los servicios que cada agente del SMA ofrece (ver Figura 5.12).

En el siguiente apartado se describe un caso de estudio en particular donde se lleva a cabo el análisis de series de tiempo para extraer de reglas de asociación basados en correlación.

6.2 Descripción del caso de estudio

En la presente tesis, se le ha dado relevancia al tópico de reglas de asociación, siendo de importancia el hecho de generar dichas reglas; en capítulos anteriores se menciona incluso un ejemplo, describiendo la metodología para generar dichas reglas.

Nuestro caso de estudio tiene el propósito de implementar un prototipo basado en la tecnología de agentes que permita la generación de reglas de asociación, considerando la información de series de tiempo de pozos petroleros, información proporcio-

nada por el Instituto Mexicano del Petróleo (IMP), donde se lleva a cabo el análisis de series de tiempo de producción de aceite, gas y agua.

Esta información consiste en series de tiempo cuya recopilación pertenece a la producción de pozos petroleros, información recabada desde el año de 1985 a diciembre del 2004.

En el capítulo anterior se mencionan algunos aspectos que deberían tomarse en cuenta para el análisis y diseño del SMA, pero en este apartado se describen las partes que fueron implementadas y los resultados obtenidos, entre los cuales el prototipo implementado realiza los siguientes pasos:

1. El sistema lleva a cabo la lectura de datos (es decir series de tiempo), estos datos son seleccionados por el usuario.

2. El sistema presenta de manera gráfica las series de tiempo seleccionadas por parte del usuario.

3. El sistema permite al usuario establecer la condición de las reglas a generar, mediante valores lingüísticos (por ejemplo: una producción alta, media y baja), también el sistema permite seleccionar una serie de tiempo base, que es aquella que nos ayudará a obtener asociación entre las reglas consideradas en el análisis a partir de la selección de fragmentos de las series de tiempo que fueron seleccionados por el usuario.

4. El sistema realiza los cálculos estadísticos correspondientes, los cuales son importantes para obtener una matriz de coeficientes de correlación; dicha matriz nos indica la asociación entre los valores que se están analizando que en la siguiente sección se interpretan los resultados de dicha matriz.

5. Por último el sistema obtiene las reglas de asociación a partir de la matriz de correlación obtenida en el punto anterior.

6.3 Resultados Experimentales

El análisis de los datos de producción de los pozos, se hace aplicando la técnica estadística de correlación lineal. Esta técnica relaciona los conjuntos de datos obtenidos experimentalmente, encuentra y describe las relaciones causales de tipo lineal que pueden darse entre los conjuntos como pares.

En el presente caso de estudio se aplica esta técnica para encontrar las posibles relaciones entre los datos de producción de los pozos. El grado de relación y el tipo

Valor Lingüístico
Alto
Medio
Bajo

Tabla 6.1: *Valores Lingüísticos*

de relación que se establece entre las variables se determina al observar el valor y el signo del *coeficiente de correlación*, definido en (3.1). En un acercamiento directo se puede considerar que una muy buena correlación entre dos variables se da cuando el valor del coeficiente de correlación se acerca a $+1$ o -1. Cuando el valor del factor de correlación es cercano a cero, la correlación es mala.

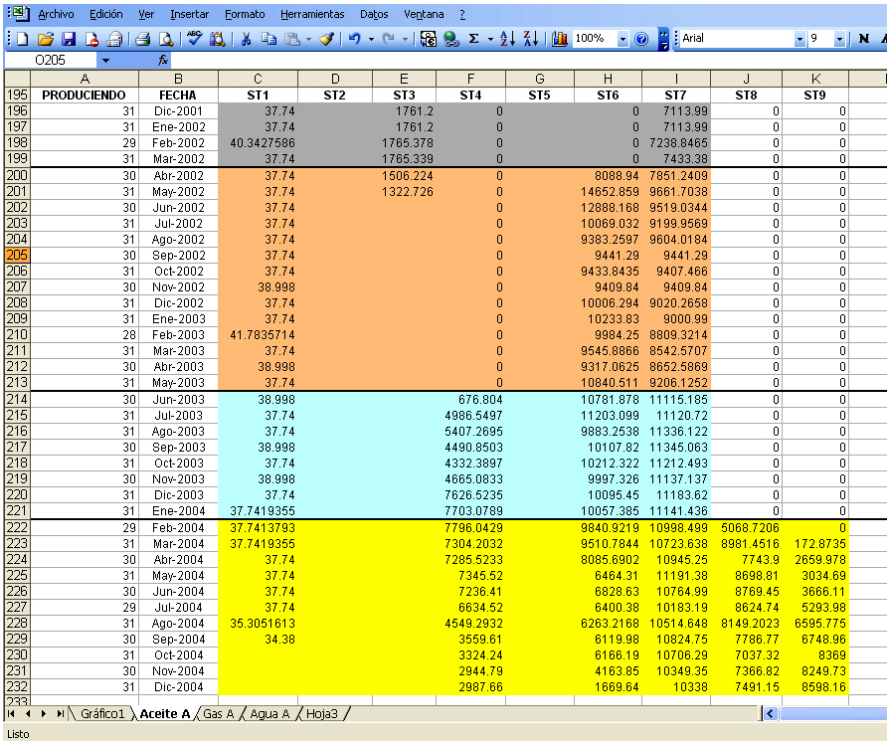
Sin embargo, una correlación fuerte, positiva o negativa no siempre indica una relación causa-efecto. Por ejemplo, si entre dos variables se espera una correlación positiva y el análisis indica una correlación negativa esto puede significar que existe otro factor que afecta los datos y que, posiblemente, no esté considerado en el análisis [Canavos,1988].

Considerando la metodología descrita en el capítulo 4 y la información proporcionada por el IMP, se llevo a cabo el análisis de series de tiempo de la siguiente manera:

Como primer punto se hace la lectura de los datos, en nuestro caso de estudio, consideramos las series de tiempo de nueve pozos petroleros (*Datos de Entrada*) y la elección de la variable, donde únicamente consideramos la producción de Aceite (ver Figura 6.1). Las series de tiempo elegidas fueron: ST6, ST7, ST8 y ST9.

El siguiente paso es establecer los valores lingüísticos y la selección de una serie de tiempo base. En la Tabla 6.1, se muestran los valores lingüísticos permitidos en nuestra aplicación. Por ejemplo, podemos seleccionar el valor lingüístico "Alto", en caso de seleccionar "Alto" debemos tomar en cuenta que valor máximo es el que queremos considerar, en tal caso podemos definir el valor máximo igual a 7000.

Es importante tomar en cuenta la serie de tiempo base, en la Tabla 6.2, se muestran las series de tiempo que fueron ingresados en la etapa de selección. Por ejemplo podemos considerar la ST8, para posteriormente hacer los cálculos estadísticos correspondientes. Estos se llevan a cabo formando una matriz de datos, la cual es muy importante para cualquier procedimiento estadístico. La Figura 6.2, se muestra las series de tiempo que han sido seleccionadas y presentadas por el Agente asistente personal del usuario.



	A	B	C	D	E	F	G	H	I	J	K
195	PRODUCIENDO	FECHA	ST1	ST2	ST3	ST4	ST5	ST6	ST7	ST8	ST9
196	31	Dic-2001	37.74		1761.2	0		0	7113.99	0	0
197	31	Ene-2002	37.74		1761.2	0		0	7113.99	0	0
198	29	Feb-2002	40.3427586		1765.378	0		0	7238.8465	0	0
199	31	Mar-2002	37.74		1765.339	0		0	7433.38	0	0
200	30	Abr-2002	37.74		1506.224	0		8088.94	7851.2409	0	0
201	31	May-2002	37.74		1322.726	0		14652.859	9661.7038	0	0
202	30	Jun-2002	37.74			0		12888.168	9519.0344	0	0
203	31	Jul-2002	37.74			0		10069.032	9199.9569	0	0
204	31	Ago-2002	37.74			0		9383.2597	9604.0184	0	0
205	30	Sep-2002	37.74			0		9441.29	9441.29	0	0
206	31	Oct-2002	37.74			0		9433.8435	9407.466	0	0
207	30	Nov-2002	38.998			0		9409.84	9409.84	0	0
208	31	Dic-2002	37.74			0		10006.294	9020.2658	0	0
209	31	Ene-2003	37.74			0		10233.83	9000.99	0	0
210	28	Feb-2003	41.7835714			0		9984.25	8809.3214	0	0
211	31	Mar-2003	37.74			0		9545.8866	8542.5707	0	0
212	30	Abr-2003	38.998			0		9317.0625	8652.5869	0	0
213	31	May-2003	37.74			0		10840.511	9206.1252	0	0
214	30	Jun-2003	38.998			676.804		10781.878	11115.185	0	0
215	31	Jul-2003	37.74			4986.5497		11203.099	11120.72	0	0
216	31	Ago-2003	37.74			5407.2695		9883.2538	11336.122	0	0
217	30	Sep-2003	38.998			4490.8503		10107.82	11345.063	0	0
218	31	Oct-2003	37.74			4332.3897		10212.322	11212.493	0	0
219	30	Nov-2003	38.998			4665.0833		9997.326	11137.137	0	0
220	31	Dic-2003	37.74			7626.5235		10095.45	11183.62	0	0
221	31	Ene-2004	37.7419355			7703.0789		10057.385	11141.436	0	0
222	29	Feb-2004	37.7413793			7796.0429		9840.9219	10998.499	5068.7206	0
223	31	Mar-2004	37.7419355			7304.2032		9510.7844	10723.638	8981.4516	172.8735
224	30	Abr-2004	37.74			7285.5233		8085.6902	10945.25	7743.9	2659.978
225	31	May-2004	37.74			7345.52		6464.31	11191.38	8698.81	3034.69
226	30	Jun-2004	37.74			7236.41		6828.63	10764.99	8769.45	3666.11
227	29	Jul-2004	37.74			6634.52		6400.38	10183.19	8624.74	5293.98
228	31	Ago-2004	35.3051613			4549.2932		6263.2168	10514.648	8149.2023	6595.775
229	30	Sep-2004	34.38			3559.61		6119.98	10824.75	7786.77	6748.96
230	31	Oct-2004				3324.24		6166.19	10706.29	7037.32	8369
231	30	Nov-2004				2944.79		4163.85	10349.35	7366.82	8249.73
232	31	Dic-2004				2987.66		1669.64	10338	7491.15	8598.16

Figura 6.1: *Muestra las series de tiempo contenidos en archivos Excel*

En la Tabla 6.3 se muestran los procedimientos estadísticos realizados y los datos iniciales, los cuales se llevan a una Matriz de Datos definido en (6.1).

$$\mathbf{X} = \begin{pmatrix} 231 & 1286.3514 & 1669.64 & 14652.859 & 3287.5982 \\ 231 & 2127.1049 & 10338 & 11345.063 & 3898.8652 \\ 231 & 371.075 & 7491.1499 & 8981.4516 & 1679.0676 \\ 231 & 231.1223 & 8598.1599 & 8598.1599 & 1227.4418 \end{pmatrix} \tag{6.1}$$

Para llevar a cabo el análisis de series de tiempo y extraer reglas de asociación, se tiene que obtener una matriz de coeficientes de correlación, ésta matriz la podemos obtener aplicando la siguiente fórmula la cual permite la programación de dichos procedimientos:

$$R = S^{-1}CS^{-1} \tag{6.2}$$

donde:

S^{-1}, es la inversa de la Matriz de Covarianza C y
C, es la Matriz de Covarianza.

Series de Tiempo
ST6
ST7
ST8
ST9

Tabla 6.2: *Series de tiempo seleccionadas*

Datos	*Promedio*	*Mínimo*	*Máximo*	*Desviación Estándar*
231	1286.3514	1669.64	14652.859	3287.5982
231	2127.1049	10338	11345.063	3898.8652
231	371.075	7491.1499	8981.4516	1679.0676
231	231.1223	8598.1599	8598.1599	1227.4418

Tabla 6.3: *Datos iniciales*

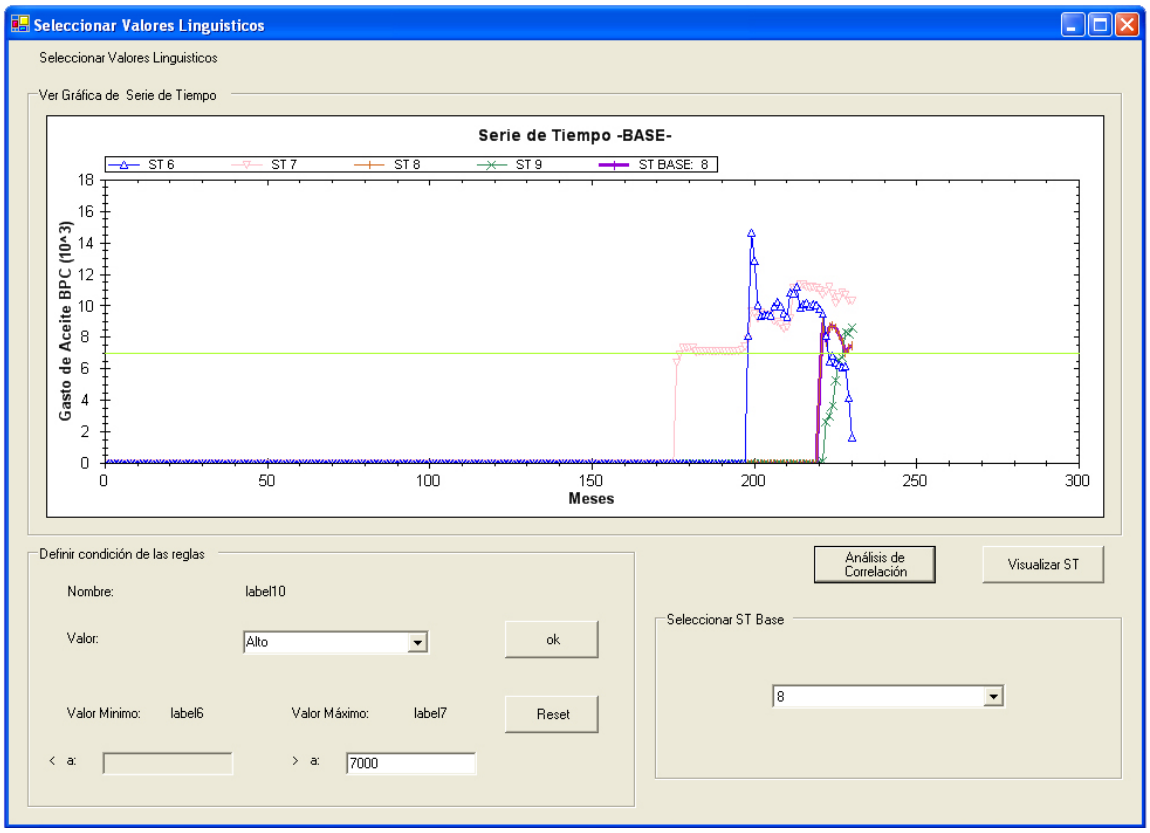


Figura 6.2: *Interfaz donde se muestra gráficamente las series de tiempo seleccionadas*

Aplicando la siguiente fórmula, obtenemos la matriz de covarianza:

$$C = CSSCP\left(\frac{1}{N-1}\right) = D^t D\left(\frac{1}{N-1}\right) \tag{6.3}$$

donde:

$CSSCP$, es la matriz resultante al obtener la suma correcta de los cuadrados y el producto de la matriz definida en (6.1), cuya diagonal principal es la suma de cuadrados y los triangulares es el producto cruz de dicha matriz.

D, es una matriz resultante la matriz de datos (6.1), la cual se le substrae el promedio de cada columna sobre cada uno de los elementos de dicha matriz.

D^t, es la traspuesta de la matriz D

La matriz D al aplicar las fórmulas descritas con anterioridad queda:

$$\mathbf{D} = \begin{pmatrix} 0 & 282.438 & -5354.5974 & 3758.4756 & 764.3549 \\ 0 & 1123.1914 & 3313.7625 & 450.6796 & 1375.6219 \\ 0 & -632.8383 & 466.9124 & -1912.9317 & -844.1756 \\ 0 & -772.791 & 1573.9224 & -2296.2234 & -1295.8014 \end{pmatrix} \tag{6.4}$$

Al obtener la traspuesta de la matriz D (6.4) nos queda:

$$\mathbf{D^t} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 282.438 & 1123.1914 & -632.8383 & -772.791 \\ -5354.5974 & 3313.7625 & 466.9124 & 1573.9224 \\ 3758.4756 & 450.6796 & -1912.9317 & -2296.2234 \\ 764.3549 & 1375.6219 & -844.1756 & -1295.8014 \end{pmatrix} \tag{6.5}$$

Aplicando la fórmula definida en (6.3), se obtiene la matriz $CSSCP$:

$$\mathbf{CSSCP} = \begin{pmatrix} 4678041.7885 & 1395709.4906 & 9105627.229 & 6593160.7208 \\ 1395709.4906 & 84695950.4078 & -46277856.1468 & -3935951.3583 \\ 9105627.2229 & -46277856.1468 & 46522402.7421 & 16166148.7319 \\ 6593160.7208 & -3935951.3583 & 16166148.7319 & 9736616.3295 \end{pmatrix} \tag{6.6}$$

Para que finalmente obtenemos la matriz de covarianza C en base a la fórmula descrita en (6.3):

$$\mathbf{C} = \begin{pmatrix} 1169510.4471 & 348927.3726 & 2276406.8057 & 1648290.1802 \\ 348927.3726 & 21173987.6019 & -11569464.0367 & -983987.8395 \\ 227606.8057 & -11569464.0367 & 11630600.6855 & 4041537.1829 \\ 1648290.1802 & -983987.8395 & 4041537.1829 & 2434154.0823 \end{pmatrix} \tag{6.7}$$

Una vez que se obtuvo la matriz de covarianza C y de acuerdo a la fórmula definida en (6.2), hay que obtener la matriz S cuya diagonal de dicha matriz está conformado por la desviación estándar de cada uno de los elementos de la diagonal de la matriz de covarianza C (6.7); la desviación estándar es obtenido por la raíz cuadrada de cada uno de los elementos de la diagonal y el valor de los elementos de los triangulares deben estar en 0.

$$\mathbf{S} = \begin{pmatrix} 1081.439 & 0 & 0 & 0 \\ 0 & 4601.5201 & 0 & 0 \\ 0 & 0 & 3410.3666 & 0 \\ 0 & 0 & & 1560.1775 \end{pmatrix} \tag{6.8}$$

Sólo resta obtener la inversa de la matriz S y es calculado con tan sólo obtener la inversa de los elementos de la diagonal de la matriz S (6.8).

$$\mathbf{S^{-1}} = \begin{pmatrix} 0.0009 & 0 & 0 & 0 \\ 0 & 0.0002 & 0 & 0 \\ 0 & 0 & 0.0002 & 0 \\ 0 & 0 & & 0.0006 \end{pmatrix} \tag{6.9}$$

Después de hacer todos estos cálculos, obtenemos la matriz de coeficientes de correlación definido en (6.2).

$$\mathbf{R} = \begin{pmatrix} 1 & 0.0701 & 0.6172 & 0.9769 \\ 0.0701 & 0.9999 & -0.737 & -0.137 \\ 0.6172 & -0.7372 & 1 & 0.7595 \\ 0.9769 & -0.137 & 0.7595 & 0.9999 \end{pmatrix} \tag{6.10}$$

Finalmente se extraen las siguientes reglas de asociación a partir de la matriz de coeficientes de correlación (6.10):

If "producción" ST8 es "Alto" Then:

$ST6 \rightarrow ST7$ isr P(ST6,ST7)=0.0701
$ST6 \rightarrow ST8$ isr P(ST6,ST8)=0.6172
$ST6 \rightarrow ST9$ isr P(ST6,ST9)=0.9769
$ST7 \rightarrow ST8$ isr P(ST7,ST8)=-0.7372
$ST7 \rightarrow ST9$ isr P(ST7,ST9)=-0.137
$ST8 \rightarrow ST9$ isr P(ST8,ST9)=0.7595

Estas son otras reglas generadas bajo las siguientes condiciones, las series de tiempo seleccionadas fueron: ST2, ST5, ST7 y ST9. La serie de tiempo base fue ST7 y el valor lingüístico fue igual a 7000.

If "producción" ST7 es "Alto" Then:

$ST2 \rightarrow ST5$ isr P(ST2,ST5)=0.6616
$ST2 \rightarrow ST7$ isr P(ST2,ST7)=0.7478
$ST2 \rightarrow ST9$ isr P(ST2,ST9)=0.981
$ST5 \rightarrow ST7$ isr P(ST5,ST7)=0.9925
$ST5 \rightarrow ST9$ isr P(ST5,ST9)=0.7539
$ST7 \rightarrow ST9$ isr P(ST7,ST9)=0.8619

En este último ejemplo las series de tiempo seleccionadas fueron: ST1, ST2, ST3 y ST4. La serie de tiempo base fue ST3 y el valor lingüístico fue igual a 3000.

If "producción" ST3 es "Bajo" Then:

$ST1 \rightarrow ST2$ isr P(ST1,ST2)=-0.0232
$ST1 \rightarrow ST3$ isr P(ST1,ST3)=0.1025

$ST1 \rightarrow ST4$ isr P(ST1,ST4)=0.3723
$ST2 \rightarrow ST3$ isr P(ST2,ST3)=0.9908
$ST2 \rightarrow ST4$ isr P(ST2,ST4)=0.9119
$ST3 \rightarrow ST4$ isr P(ST3,ST4)=0.9587

En base a las reglas de asociación extraídas, en la Figura 6.3, se presenta gráficamente la ubicación de los pozos y los resultados de asociaciones entre ellos.

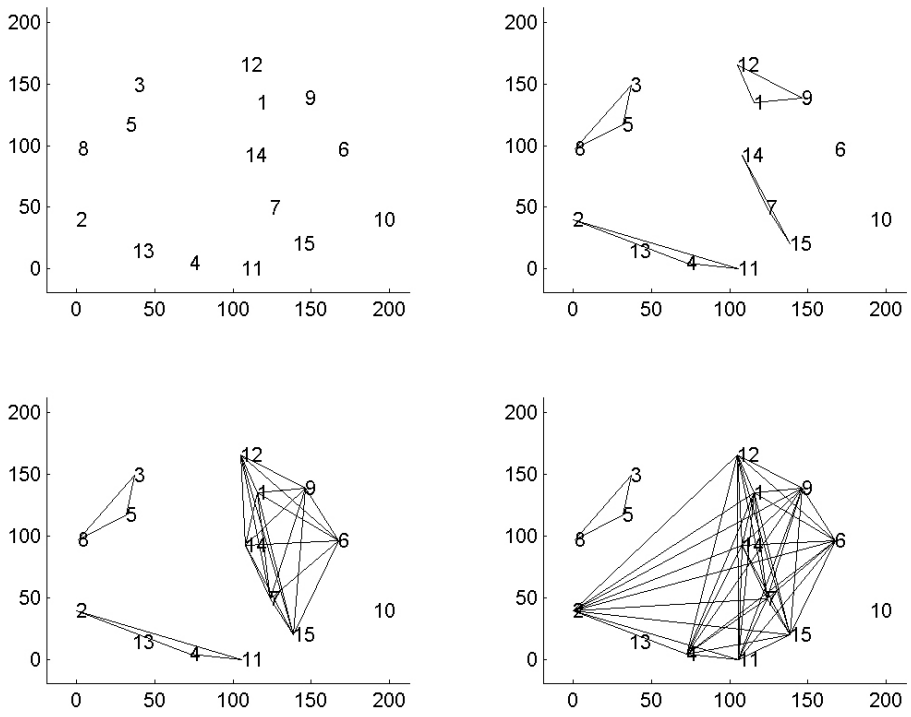


Figura 6.3: *Visualización de resultados de asociaciones entre los pozos*

Capítulo 7

Conclusiones

En el presente Capítulo se describen los resultados obtenidos, las aportaciones del desarrollo de esta tesis, así como las conclusiones del presente trabajo respecto a los objetivos planteados, terminando con el trabajo futuro propuesto de la presente tesis.

7.1 Resultados

En función a los objetivos establecidos [Capítulo 1] para el desarrollo de la esta tesis, se presenta el estado de arte donde se hace mención de aquellos sistemas que extraen reglas de asociación a partir de diversas fuentes de datos, se mencionan también aquellos sistemas que hacen uso del paradigma de agentes para generar reglas de asociación considerando las plataformas de agentes existentes en el mercado, incluso CAPNET, plataforma usada para el desarrollo de la presente tesis. [Capítulo 2]; posteriormente se llevó a cabo una revisión de aquellos conceptos teóricos que se involucraron para el desarrollo de la presente tesis [Capítulo 3]. Se propuso una metodología para generar reglas de asociación a partir de series de tiempo [Capítulo 4]. Se presenta una metodología para el desarrollo del prototipo implementado utilizando la tecnología de SMA, se hace un tratado del análisis y diseño del SMA, asi como de la arquitéctura del sistema[Capítulo 5].

Es importante hacer mención de otras metodologías que han sido desarrolladas para la extracción de reglas de asociación, incluso el mismo sistema experto Smart-Agua, utiliza un método de reconocimiento de patrones basado en la transformada de aproximación de movimiento (MAP); el cual está implementado como parte del *toolbox de Percept-Miner* [Batyrshin & Sheremetov,2005]. Cuya metodología consiste en que las series de tiempo muestran el comportamiento de la producción de agua durante cierto período; el patrón objetivo es representado como una secuencia de valores de la pendiente de la curva $(c_1, \ldots c_k)$, la transformada MAP reemplaza valores de la

serie de tiempo y de producción de agua con la secuencia valores de la pendiente $MAP(y) = (a_1, \ldots, a_N)$ de las líneas $f = a_i t + b$, aproximando los datos de las series de tiempo con una ventana deslizable. Basándose en la medida de distancia entre secuencias de valores de la pendiente, el método busca el patrón $(a_i, \ldots, a_{i+k-1})$ en MAP(y) más cercano (found) al patrón indicado. De esta forma se contrasta la evolución de los datos reales contra cuatro patrones típicos y relacionados a la presencia de ciertos problemas [Yortsos et al,1997]. La herramienta permite detectar con mayor confiabilidad el patrón de comportamiento en la producción de agua. La cercanía encontrada con respecto a cada tipo de curva, representa valiosa información tanto para el usuario como para el sistema durante el proceso de diagnóstico. Es importante tomar en cuenta que el modulo de minería de datos, no sólo permite identificar patrones, también permite identificar relaciones entre pozos y construir reglas de asociación.

El método llamado Box-Jenking propuesto en [Luciv & Novikov,2005], tiene como objetivo primordial la obtención de reglas de asociación en secuencias temporales, siendo su tarea principal el encontrar una fórmula matemática que describa la dependencia de las concurrencias del valor de las series de tiempo en contraste con valores previos. Si se tiene más de una serie de tiempo, éstos influyen en ellas cuyos valores pueden ser incluidos en la fórmula. Box-Jenking está basado en la teoría estadística que es usada en la predicción en series de tiempo, donde se consideran patrones A, B asociados a una serie de tiempo temporal predictivo S_p y una secuencia temporal a considerar S, seleccionado cualquier punto en el tiempo (t) de manera aleatoria, se considera que la ocurrencia de (B, t) es causado por (A, t), donde la probabilidad de dichas ocurrencias son asociadas a algun tipo de densidades (D) de ocurrencias del patrón B. Si $|D_{B,A} - D_B| > \varepsilon$, donde $\varepsilon > 0$, entonces suponemos que la ocurrencia del patron A en S_p se puede inferir de la probabilidad de la ocurrencia del patrón B en S. Dependiendo del signo en esta diferencia, se puede generar cualquiera de estas dos reglas:

$D_{B,A} - D_B > 0 : A$ ocurre $\Rightarrow B$ ocurre
$D_{B,A} - D_B < 0 : A$ ocurre $\Rightarrow B$ no ocurre

Donde las reglas obtenidas, dependerá de la probabilidad de dichas ocurrencias en el tiempo.

En resumen se muestra una tabla comparativa (ver Tabla 7.1) de las metodologías descritas con aterioridad, respecto a la metodología descrita en la presente tesis.

Metodología	*Desarrolladores*	*Tipo de Análisis*	*SMA*	*Tipo de Reglas de Asociación*
Metodología Pruesta	CIC-IPN	Coeficientes de Correlación	si	Basadas en Correlación
Minería de ST Perceptual	IMP	Percept-Miner	si	Basadas en Patrones
Box-Jenking	Universidad de San Petersburgo	Probabilidad Estadística	no	Basadas en Secuencias Temporales
F-APACS	Politécnica de Hong Kong	Diferencia de Ajuste	no	Basadas en Conjuntos Difusos

Tabla 7.1: *Tabla comparativa de las metodologías*

7.2 Contribuciones

De acuerdo a los resultados obtenidos podemos establecer las siguientes contribuciones:

- Una metodología para extraer reglas de asociación basadas en correlación a partir del análisis de series de tiempo.
- Una arquitectura multiagente que utiliza un modelo de servicios, donde la metodología descrita en el punto anterior es aplicada en los servicios que cada agente ofrece.
- Un prototipo que implementa tanto la metodología, como la arquitectura multiagente propuestos en la presente tesis para la extracción de reglas de asociación basados en correlación.

7.3 Conclusiones

La extracción de reglas de asociación ha sido uno de los tópicos en donde la minería de datos ha tomado importancia en los últimos años; su aplicación en la industria del

aceite y gas ha tomado interés por el hecho de apoyar al procesamiento de grandes volúmenes de información.

En la presente tesis, se propuso y se implementó una metodología que permitiera el análisis de series de tiempo para extraer reglas de asociación basadas en correlación aplicando la tecnología de sistemas multi-agentes cuya arquitectura, está basada en la solicitud y prestación de servicios por diferentes entidades participantes en un ambiente que es de objeto de estudio, donde los agentes registran sus servicios en el DF de la plataforma CAPNET, con el propósito de que dichos servicios sean accesibles para los agentes que lo soliciten.

7.4 Trabajo Futuro

De acuerdo al trabajo desarrollado en la presente tesis, se pueden planear los siguientes objetivos a futuro:

- Extender la metodología propuesta en la presente tesis para extraer reglas de asociación a partir de series de tiempo considerando valores difusos.
- En base al punto anterior, implementar el prototipo que permita extraer reglas de asociación a partir de valores difusos.

Bibliografía

[Agent Builder,2002] Agent Builder,(2002) *An integrate Toolkit for Constructing Intelligent Software Agents*. Guía de usuario , version 1.3 , 30 de abril del 2002. Reticular System, Inc. `http://www.agentbuilder.com`

[Agrawal & Arning et al,1996] Agrawal ,Rakesh., Mehta, Manish., Shafer, John. y Srikant, Ramakrishnan. Arning, Andreas y Toni, Bollinger (1996) *The Quest Data Mining System*. IBM Almaden Research Center San Jose, California, U.S.A. y IBM German Software Development Laboratory Boeblingen, Germany

[Bailey et al,2000] Bailey B., Crabtree M. et al. *Control de agua*. Oilfield Review 2000 (Schlumberger).

[Batyrshin et al,2004] Batyrshin I., Herrera-Avelar R. ,Sheremetov L., Suarez R. *Mining fuzzy association rules and networks in time series databases*, Proc. Int. Conf. Fuzzy Sets Soft Computing in Economics and Finance, FSSCEF 2004, St. Petersburg, Russia, 2004, vol. I, 39-53.

[Batyrshin & Sheremetov,2005] Batyrshin I., Sheremetov L., *Perception Based Time Series Data Mining with MAP Transform*, MICAI-2005, Advances in Artificial Intelligence, LNAI. Springer Verlag. 2005

[Batyrshin & Sheremetov,2006] Batyrshin I.Z, Sheremetov L.B., *Perception based associations in time series data bases*. NAFIPS 2006, Montreal, 2006.

[Bigus & Bigus,2001] Joseph P. Bigus, Jennifer Bigus. (2001), *Contructing Intelligent Agents Using Java*, Second Edition, John Wiley and Sons Inc. 2001.

[Botía et al,2005] Juan Antonio Botia Blaya, Juan Manuel Hernansáez Amor y Antonio F. Gómez-Skarmeta. *Asistencia Personalizada a la Minería de Datos mediante Agentes Inteligentes*. Departamento de Ingeniería de la Información y las Comunicaciones. Universidad de Murcia. 21 de Marzo de 2005.

[Blum & Furst,1997] Blum, L. y Furst, L. (1997) *Fast Planning Through Planning Graph Analysis* Tomado de la Artificial Intelligence School Of Computer Science Cargenie Mellon University Pittsburgh.

[Brin et al,1997] Brin S., Motwani R. y Silverstein C. (1997) *Beyond Market Baskets: Generalizing Association Rules to Correlations*, Proceedings of the ACM SIGMOD/PODS '97 Joint Conference. May 1997, pp. 265-276.

[Canavos,1988] Canavos, G.C. (1988) *Probabilidad y estadística. Aplicaciones y método.* Editorial McGraw-Hill/Interamericana de México SA de CV, México 1988.

[Chan & Au,1997] Keith C. C., Chan. y Wai-Ho, Au.(1997) *Mining Fuzzy Asociation Rules* Deparment of Computing, The Hong Kong Polytechnic University. Hung Hom, Kowloon, Hong Kong. Tomado de la Biblioteca Digital de ACM. Enero 1997 Proceedings of the sixth international conference on Information and knowledge management CIKM '97 Publisher: ACM Press

[Chou,1977] Ya-Lun Chou. *Análisis Estadístico* 2da Edición. Ed. Iberoamericana 1977 México.

[Contreras et al,2004] Miguel Contreras, Ernesto Germán, Manuel Chi y Leonid Sheremetov. *Design and implementation of a FIPA compliant agent platform in .NET*. Publicado en el Journal of Object Tecnology.. por ETH Zurich, Chair of software engineering 2004 Vol 3. No.9

[De Cock et al,2005] De Cock M., Cornelis C., Kerre E.E., *Elicitation of fuzzy association rules from positive and negative examples, Fuzzy Sets and Systems* 149 (1) (2005) 73-85.

[Dubois et al,2005] Dubois D., Prade H., Sudkamp T., *On the representation, measurement, and discovery of fuzzy associations*, IEEE Trans. Fuzzy Systems, 13 (2005) 250-262.

[Ferber & Gutknecht,1998] J. Ferber y O. Gutknecht. *A meta-model for analysis and design of multi-agent systems*, Proceedings of the 3rd International Conference on Multi-Agent Systems, (ICMAS98), IEEE, pp. 155-176, August 1998.

[Goebel & Gruenwald,1999] Goebel, Michael y Gruenwald, Le *A survey of data mining and knowledge discovery software tools* ACM SIGKDD, Vol 1 Issue 1. JUNE 1999. Tomado de la biblioteca digital de ACM.

[FIPA,www] Foundation for Intelligent Physical Agents (FIPA), `http://www.fipa.org/about/index.html`

[Han & Kamber,2001] Jiaewi, Han y Micheline, Kamber. (2001) *Data Mining: Concepts and Techniques*, Simon Fraser University. 2001 Academic Press. USA.

[Hand et al,2001] David, Han. Heikki, Mannila y Padhraic, Smyth. (2001) *Principles of Data Mining*, A Bradford Book. The MIT Press Cambridge, Massachusetts London, England.

[Hannebauer & Geske,1999] Hannebauer , M. y Geske, Ulrich. (1999) *Coordination Distributed CLP-Solvers in Medical Appointment Scheduling*, Planning and Optimization Laboratory.

[Heecheol et al,2000] Heecheol, J. Petrie, C. y Cutkosky M. R. *JATLite: A Java Agent Infraestructure with Message Routing*.

[Hietala et al,1999] Hietala, P. Y Niemirepo, T. (1999) *Studying learner - computer interaction in agent bases social learning environments*, University of Tampere, Tampere Finland, 1999

[JADE,www] JADE *Java Agent Development Framework*. `http://sharon.cselt.it/projects/jade/`

[JATLiteBeta,www] *JATLite Beta Complete Documentation* (1998), Stanford University, `http://java.stanford.edu`

[Jeannings & Wooldrigde,1998] Jeannings, N. Y Wooldridge M.(1998) *Agent Technology Foundations, Applications, and Markets*. Ed. Springer 1998.

[Johnson,1998] Dallas E. Johnson (1998) *Métodos multivariados aplicados al análisis de datos* International Thomson Editores 2000. Traducción del libro Applied Multivariate Methods for Data Analisys publicado en inglés por Brooks Cole 1998.

[Keith & Whai,1997] Keith C.C. Chan, Whai-Ho Au (1997) *Mining Fuzzy Association Rules*. CIKM'97 Las Vegas Nevada, USA. Tomado de la Biblioteca de ACM. ACM 0-89791-970-x/97 11

[Klijin,2001] Flip Klijin, Abril 2001 *Análisis Cluster - Análisis de Conglomerados (AC): Taxonomía Numérica*.

[Klusch, Lodi & Moro,2003] [Klusch, Matthias Klush, Stefano Lodi y Gianluca Moro.(2003) *Agent-Based Distributed Data Mining: The KDEC Scheme* AgentLink, volume 2586 of LNCS.Singer 2003.

[Klusch, Lodi & Moro,Julio2003] Klush, Stefano Lodi y Gianluca Moro. *Issues of Agent-Based Distributed Data Mining*. AAMAS'03, July 14-18 2003 Melbourne, Australia.

[Leonid et al,2006] Leonid Sheremetov, Ildar Bathyrsin, Ana Cosultchi, Jorge Martínez-Munoz.(2006) *SMART-Agua: a Hybrid Intelligent System for Diagnostics*. INES 2006 10th IEEE International Conference on Intelligent Engineering System; London ;UK on June 26.-28, 2006

[Levin & Rubin,2004] Richard I. Levin y David S. Rubin. *Estadística para Administración y Economia* Septima edición. Ed. Pearson. 2004 México.

[Luciv & Novikov,2005] Luciv, Elena y Novikov, B. (2005) *Discovery of Association Rules in Temporal Sequences*. Proceedings of the Spring Young Researcher's Colloquium on Database and Information Systems (SYRCoDIS'2005). Saint-Petersburg State University. `http://syrcodis.citforum.ru/2005/content_en.shtml`

[MADKIT,www] MADKIT, *A Multi Agent Development, kit.* http://www.madkit.org

[Michie et al,1994] D. Michie, D.J. Spiegelhalter, C.C. Taylor. *Machine Learning, Neural and Staticstis Classification*. Ed. D. Michie, D.J. Spiegelhalter, C.C. Taylor. Febrero 1994.

[Montoro,www] *Apuntes de Métodos estadísticos en la ingeniería*. Delia Montoro Carzola. Grupo de Investigación en Estadística Teórica y Aplicada e Investigación Operativa. Universidad de Jaén, España. Tomado de la página `http://estio.ujaen.es/Asignaturas/EUP/metodos/apuntes.pdf`

[Nwana & Ndumu,1998] Nwana, H. Y Ndumu, D. (1998) *A Brief Introduction to Software Agent Technology.*

[O'Brien & Nicol,1998] PD O'Brien , R C Nicol. (1998), *Towards a Standar for Software Agents*, Tomado de la página de FIPA.

[Reyes & García,2005] José Fernando Reyes Saldaña y Rodolfo García Flores. (2005), *El proceso de descubrimiento de conocimiento en bases de datos.*, Ingenierías, Enero -Marzo 2005, Vol. VIII, No. 26, Tomado de la página `http://ingenierias.uanl.mx`

[Shahab,2005] Shahab D. Mohaghegh. *Recent Developments in Application of Artificial Intelligence in Petroleum Engineering.* J. of Petroleum Technology, 2005, pp 86-91.

[Shoham,1993] Shoham, Y. *Agent Oriented Programming, Artificial Intelligence*, 6: pp. 51-92, 1993.

[Winton & Pete,1995] Winton H. E. Davies y Pete Edwards.(1995) *Agent - Based Knowledge Discovery* Departmet of Computing Science. King´s College, University of Aberdeen. Uk. In AAAI Spring Symposium on Information Gathering, AAAI Press, 1995. `http://www.agent.ai/main.php?folderID=211`

[Wooldrigde & Jeannings,1994] Wooldridge M. y Jeannings, N. (1994) *Intelligent Agents: Theory and Practice*. Registrado en Knowledge Engineering Review , Octubre 1994, Revisado en 1995.

[Wu & Soibelman,2005] Jianfeng Wu y Lucio Soibelman. (2005)*Data Fusion and Knowledge Discovery from Construction Scheduling Data*. 2005 AIS Symposium, Department of Civil and Environmental Engineering Advanced Infrastructure Systems, Carnegie Mellon University. `http://www.ce.cmu.edu/ais/research.html`

[Yairi et al,2001] Takehisa, Yairi., Yoshikiyo, Kato y Koichi, Hori (2001) *Fault Detection by Mining Association Rules from House-keeping Data*. RCAST, University of Tokyo. Tokyo, JAPAN. In Proc. of International Symposium on Artificial Intelligence, Robotics and Automation in Space, 2001. Tomado de la Biblioteca Digital de ACM.

[Yortsos et al,1997] Yortsos, Y.C., Choi, Y., Yang, Z. and Shah, P.C., *Analysis and Interpretation of Water/Oil Ratio in Waterfloods*, SPE 59477

[Zeus,1999] *The ZEUS Technical Manual*, Colis J., Ndumu D., external documentation, `http://labs.bt.com/projects/agents/zeux/` ,September 1999.

[Zhang,2003] Zili Zhang *An Agent-Based Hybrid Framework for Database Mining* Faculty of Information Technology, UTS, Sydney , Australia. Publicado: Artificial Intelligence. 2003 Taylos Francis.

[Zhang & Zhang,2002] Chengqi Zhang y Shichao Zhang *Association Rule Mining, Models and Algorithms* University of Technology, Sydney, Faculty of Information Technology. Sydney Australia. Editorial Springer 2002.