ISSN: 1692-7257 Volumen 1 - 2006

MULTI-AGENT SYSTEM FOR DIGITAL DOCUMENTS RECUPERATION

SISTEMA MULTIAGENTE PARA LA RECUPERACIÓN DE DOCUMENTOS DIGITALES

Jaime Guzmán, Alejandro Sánchez, Durley Torres

Universidad Nacional de Colombia
Facultad de Minas, Escuela de Sistemas, Medellín, Colombia
{jaguzman, asanchez}@unalmed.edu.co, durleytp@gmail.com

Abstract: This paper presents the process required for the construction of a Multi-Agent system for recuperation of digital documents using the methodology of modeling Multi-Agent Systems, MAS-CommonKADS. This system is based on the Vectorial Model and it was implemented in JADE, a framework for the development of Multi-Agent systems fully implemented in Java language. The digital documents retrieval system allows to users to make queries in natural language and to enter and to process digital documents in its documental base.

Resumen: En este artículo se presenta el proceso seguido para la construcción de un sistema Multi-Agente de recuperación de documentos digitales utilizando la metodología de modelado de sistemas Multi-Agente MAS-CommonKADS. Este sistema se basa en el modelo de recuperación de información Vectorial y está implementado en la plataforma JADE, una herramienta para el desarrollo de sistemas Multi-Agente totalmente implementada en Java. Este sistema de recuperación permite consultas por parte del usuario en lenguaje natural y el ingreso y procesado de los documentos digitales que conforman la base documental del sistema.

Keywords: Information retrieval system, Agents technology, Digital documentation center, MAS-CommonKADS methodology.

1. INTRODUCCION

La recuperación de información (RI) es una disciplina de creciente interés, dado el aumento de la disponibilidad de documentos en soporte digital y de la necesidad consiguiente de obtener en cada momento aquellos que responden a una necesidad informativa dada. Asociada a esta, la teoría de agentes (Arenas, 2002; Hendler, 2004) ha entrado a jugar un papel importante en el desarrollo de esta clase de sistemas al permitir una mayor flexibilidad en su escalabilidad lo que ha facilitado un mejor manejo de la sobrecarga de información en grandes volúmenes de documentos.

El presente artículo, ilustra la construcción del sistema S.A.B.I.O. (Sistema de Almacenamiento y Búsqueda de Información Organizada), el cual permite el almacenamiento y recuperación de documentos digitales utilizando el conocido modelo del espacio vectorial (Baeza-Yates, 1999) y la tecnología de agentes para su soporte,

mediante el uso de JADE (Bellifemine et al, 2004), una plataforma para el desarrollo de sistemas Multi-Agente (Iglesias, 1998).

La necesidad de darle un carácter de ingeniería al desarrollo de sistemas Multi-Agente, ha generado la aparición de diversas metodologías para el análisis y diseño de software basado en agentes. Este trabajo presenta el ciclo de vida del sistema utilizando la metodología MAS-CommonKADS (Schreiber et al, 1994), una metodología para el desarrollo de sistemas Multi-Agente.

El presente artículo está organizado de la siguiente manera: en la siguiente sección, se describen de manera muy general las técnicas de solución usadas por el sistema S.A.B.I.O., En la sección 3 se presenta la conceptualización del mismo sistema. La sección 4 expone la fase de análisis del modelamiento del sistema. En el numeral 5, se detalla la fase de diseño del sistema S.A.B.I.O. La sección 6, presenta las conclusiones y trabajos futuros.

2. VISIÓN GENERAL DEL SISTEMA

La Recuperación de Información (RI) trata la representación, almacenamiento, organización y acceso a las unidades de información de los textos (hipertextos, multimedia) de una colección de documentos. Particularmente la RI pretende determinar qué documentos de una colección dada son relevantes para las necesidades de información de un usuario representadas en una consulta. Así se pretende proporcionar solo la información útil según los requerimientos de usuario y evitar desplegar la información no relevante (Bellifemine et al, 2004).

2.1 El modelo de recuperación del sistema

Los Sistemas de Recuperación de Información (SRI) están asociados a un modelo específico de recuperación. Aunque en la actualidad existe una gran variedad de dichos modelos, los que más se destacan por su tradición son (Salton, 1989): el modelo de Recuperación Booleano, el modelo del espacio vectorial, el modelo de recuperación probabilística y el modelo de recuperación Booleano Extendido.
En nuestro caso, el modelo de recuperación utilizado es el modelo del espacio vectorial (Salton et al, 1983). Este modelo trata los textos y consultas como vectores en un espacio multidimensional, donde típicamente cada dimensión del vector representa la frecuencia de la ocurrencia de un término en una colección de documentos (Salton, 1989). Consultas y textos son comparados usando los vectores, basándose en nuestro caso, en la correlación del coseno como medida de similitud ya que es fácil de calcular obteniéndose un número entre 0 y 1. Entre más pequeño es el ángulo entre los 2 vectores mayor similitud tendrán sus respectivos documentos y consultas. En este modelo tanto los términos de la consulta como los términos del documento pueden ser calificados con cierto peso y la similitud computacional entre las consultas y los registros almacenados hace que sea posible obtener salidas clasificadas en un orden decreciente a la similitud consulta-documento.

Algunas razones por las que se seleccionó este modelo son: su simplicidad, lo fácil que se pueden acomodar los términos de peso, y que provee resultados ordenados de manera decreciente de las similitudes consulta−documento. Sumado a lo anterior, esta técnica tiene la virtud de facilitar la modificación de los vectores siendo posible adaptar los vectores de consulta y los vectores documento a un ambiente dinámico.

2.2 El uso de agentes para la recuperación de Información

Una solución prometedora para el problema de localizar, recuperar y procesar grandes cantidades de información es el uso de agentes (FIPA, 2005; Iglesias, 1998). En este sentido el concepto de agente aplicado al desarrollo de SRI permite que las tareas de almacenamiento, clasificación, búsqueda, recuperación, filtrado e interpretación de información se optimicen y/o automaticen logrando con esto una comunicación más rápida y segura entre los usuarios y el propio sistema.

En concordancia con lo anterior, el sistema desarrollado se implementó utilizando la tecnología de agentes. Para tal fin se empleó la metodología MAS-CommonKADS, la cual es una extensión de la metodología de ingeniería del conocimiento CommonKADS (Schreiber et al, 1994) que agrega técnicas de metodologías orientadas a objetos.

En esta metodología se proponen siete modelos para el desarrollo de sistemas Multi-Agente, los cuales son: El *modelo del agente*, el cual, especifica las características de los agentes del sistema: sus capacidades de razonamiento,

habilidades, servicios, grupos de agentes a los que pertenece y clase de agentes. El *modelo de la organización*, describe la organización de la sociedad de agentes y su relación con el entorno.

El *modelo de tareas*, identifica y describe las tareas que los agentes pueden realizar. El *modelo de experiencia*, define el conocimiento necesario que deben tener los agentes para lograr sus objetivos. El *modelo de comunicación*, describe las interacciones entre un agente humano y un agente de software. El *modelo de coordinación*, ilustra las interacciones entre los agentes. Finalmente, el *modelo del diseño*, el cual mientras los otros modelos tratan el análisis del sistema Multi-agente, este modelo se utiliza para ilustrar la arquitectura y el diseño del sistema Multi-agente como paso previo a su implementación.

La aplicación de esta metodología se divide entres fases: Conceptualización, Análisis y Diseño. Dentro de estas fases se construyen los modelos anteriormente relacionados. Es así como en las secciones siguientes se detalla el sistema desarrollado a la luz de esta metodología.

3. FASE DE CONCEPTUALIZACION

El primer paso para el desarrollo de S.A.B.I.O., ha sido la fase de conceptualización que consiste en la definición y alcance de sistema, lo que involucra todo un proceso de captura de requisitos del usuario o usuarios que intervienen en el sistema; por tal razón la primera presentación consiste en un diagrama de casos de uso, que identifica de manera clara los procesos que se llevan a cabo, identificando las interacciones de los usuarios externos con el sistema y define su alcance funcional.

En S.A.B.I.O. se identificaron de forma general dos tipos de usuarios:

- *Usuario*: son los usuarios finales del sistema, cuyo interés se centra en el proceso de consulta de documentos, busca información específica.
- *Usuario Especializado*: quien es el encargado de subir los documentos digitales que van a ser analizados por el sistema, Asume el rol del bibliotecólogo o documentalista.

En la Figura 1 se puede ver el caso de uso identificado para el actor usuario.

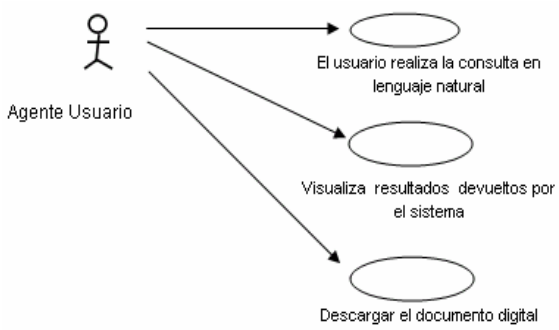

Fig. 1. Casos de uso del actor usuario

4. FASE DE ANÁLISIS

En esta fase se modela el sistema Multi-Agente propuesto en este documento, para lo cual se desarrollan uno a uno los modelos que define la metodología MAS-CommonKADS.

4.1 Modelo de Agentes

El sistema Multi-Agente con el cual se van a representar los procesos que se deben llevar a cabo en S.A.B.I.O., consta de la identificación de los siguientes agentes:

- *Agentes Humanos*: Estos agentes son agentes de interfaz, que ayudan a gestionar los procesos realizados por cada uno de los actores del sistema. Los agentes de interfaz identificados fueron: el *Agente Usuario* y el *Agente Usuario Especializado*, el primero se encarga de procesar las peticiones funcionales del usuario, es decir, las peticiones de consulta de documentos y el segundo las peticiones del usuario especializado, es decir, las peticiones de indexación de documentos digitales.
- *Agente Indexador: S*e encarga de realizar el procesamiento de los documentos incluidos en la base de documental por el Agente Usuario Especializado. Este procesamiento esta directamente ligado al modelo de recuperación de información escogido para el sistema, en este caso el modelo vectorial.
- *Agente de Búsqueda por Recuperación*: Se encarga de realizar la traducción de la consulta ingresada por el usuario en lenguaje natural al formato usado por el método vectorial, es así como se realiza: La eliminación tanto de las palabras vacías (los artículos, preposiciones, etc.),

como de los signos de puntuación y se calculan los pesos de cada término de la consulta, para seguidamente asociarlos a un vector de consulta (Salton, 1988).

- *Agente Base de Datos Documental*: Se encarga de realizar el cálculo de pesos de las palabras identificadas en cada uno de los documentos teniendo como base la formula definida en el modelo vectorial (Salton, 1983).
- *Agente de Información*: Se encarga de seleccionar los documentos que satisfacen la consulta del usuario, es decir, hace una comparación entre la consulta procesada por el Agente de Búsqueda por Recuperación y los documentos indexados y procesados por los Agentes Indexador y Base de Datos Documental.

En MAS–CommonKADS la descripción de los agentes se realiza por medio de plantillas textuales de agentes, como la detallada en la tabla 1, donde se presenta la plantilla para el Agente Usuario.

Tabla 1. Plantilla textual del Agente Usuario.

Nombre: Agente Usuario
Tipo: Agente de Interfaz **Papel:** Servir como intermediario entre el usuario del sistema y el SRI, su función principal es capturar las requerimientos informativos de los usuarios y realizar los procesos necesarios para retornarle un resultado que satisfaga sus necesidades. **Servicios**: Conocer los requerimientos informativos de los usuarios, Buscar información de interés para el usuario en el SRI, Mostrar resultados de búsqueda en un formato legible y comprensible para el usuario, Descarga de material digital **Descripción:** Agente de interfaz que captura las consultas que realizar los usuarios al SRI y realiza los procesos de búsqueda necesarios para retornar resultados de interés para el usuario.

La identificación de objetivos, planes (cómo se espera cumplir con los objetivos), colaboraciones (Agentes que ayudan para cumplir con los objetivos especificados) y conocimiento (qué conocimiento es necesario para cumplir con los objetivos) de cada uno de los agentes, se lleva a cabo por medio de las tarjetas para descripción de agentes. En la tabla 2, a manera de ejemplo, se presenta la tarjeta de descripción del Agente Usuario.

Tabla 2. Tarjeta de descripción del Agente Usuario

OBJETIVOS	PLANES	CONOCIMIENTOS	COLABORADOR	SERVICIO
Captura de las consultas de los usuarios	Brindar al usuario un formulario donde pueda expresar las necesidades informativas	Ninguno	ninguno	Conocer los requerimientos informativos de los usuarios
Buscar información en el SRI de acuerdo a los requerimientos del usuario	comunicarse con el Agente de Información para que realice procesos de búsqueda en el SRI	Ubicación del agente de información	Agente de Información	Buscar información de interés para el usuario en el SRI
Presentar los resultados de las búsquedas a los usuarios	Listar los documentos encontrados de forma comprensible al usuario	Ninguno	Ninguno	Mostrar resultados de la búsqueda en un formato claro para el usuario
Permitir la descarga de material digital	darle la opción al usuario de la descarga de material digital	Ninguno	Ninguno	Descarga de material digital

4.2 El Modelo de Tareas

En este modelo se presenta el *diagrama de actividades,* que muestra los aspectos dinámicos del sistema, el cual permite modelar los pasos secuénciales (y posiblemente concurrentes) del proceso computacional seguido por el sistema.

Una de las tareas identificadas en el sistema S.A.B.I.O. es buscar información que cumpla con los requerimientos del usuario. En este caso los agentes que intervienen para alcanzar el proceso,

- Agente Usuario: Ingresa su demanda de información.
- Agente de Búsqueda por Recuperación: Quien traslada la consulta a un formato tipo vector con los términos y pesos.
- Agente de Información: Realiza los procesos de búsqueda de la información que resulte de interés al Agente Usuario.

En el diagrama de actividad de la figura 2, se definen las tareas que debe ejecutar el sistema para buscar la información deseada por el usuario

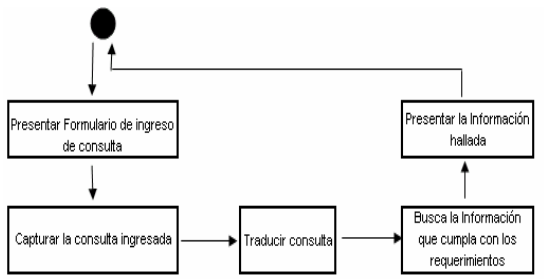

Fig. 2. Diagrama: Tarea buscar Información

4.3 Modelo de la Organización

El objetivo del modelo de la organización Multi-Agente es representar las relaciones "estáticas" entre los agentes del sistema. Para crear este modelo se llevan a cabo dos tareas: Identificación de la estructura organizacional e identificación de las relaciones de herencia

En el caso de S.A.B.I.O, la organización de los agentes, puede ser vista en la figura 3. En este diagrama además de encontrar los agentes descritos en el modelo de agentes, se encuentra una aplicación que actúa como intermediaria entre el usuario y el sistema Multi-Agente, denominada *Gateway,* que corresponde a una aplicación JSP (Tremblett, 2002) (java Server Page), que se encarga de procesar el requerimiento funcional del usuario del sistema para luego comunicarse a través de protocolos de red con el sistema Multi-Agente, generando una relación estática adicional con el modelo de agentes aquí descrito.

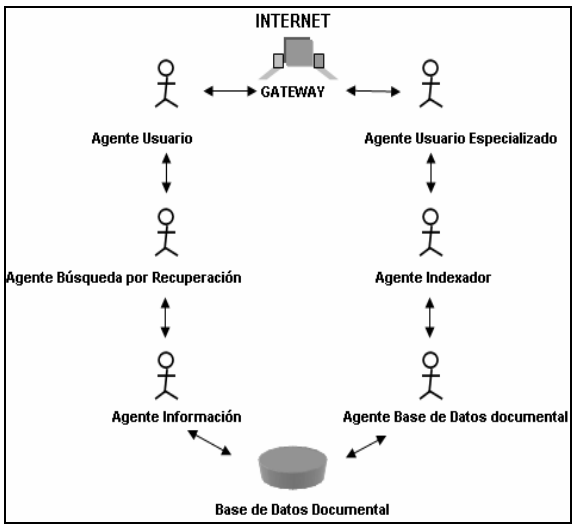

Fig. 3. Modelo de Organización

La relación de herencia entre agentes permite agrupar clases de los agentes que comparten capacidades parecidas (servicios, objetivos, etc.).

El sistema Multi-Agente de S.A.B.I.O., ilustrado en la tabla 3, propone una forma de tal agrupación:

- Tipo: define la clasificación funcional del agente.
- Clase: identifica a los agentes que tienen cierto objetivo final conjunto.
- Subclase: corresponde a un agente específico del sistema S.A.B.I.O

Tabla 3. Modelo de Organización

Tipo	Clase	Subclase
Agente de Interfaz	Agentes Usuario	Agente Usuario Agente Usuario Especializado
Agente de Software	Agentes información	Agente información
	Agentes de Tareas	Agentes Indexador Agente de Búsqueda por Recuperación Agente base de datos Documental

4.4 Modelo de Coordinación y Comunicación

En este modelo se deben presentar dos actividades fundamentales:

- Describir las interacciones entre los agentes
- Identificar las conversaciones entre los agentes y describir esas conversaciones.

Para realizar la identificación de las conversaciones se ha recurrido a los resultados de las etapas anteriores. Modelos como el de tareas y el de agentes son de gran apoyo en la identificación de las conversaciones realizadas por los agentes, ya que clarifican los procesos que se llevan acabo.

Una forma de representar las conversaciones es a través de un diagrama de casos de uso interno. A manera de ejemplo en la figura 4, se presenta de forma muy general, las conversaciones que pueden darse entre el Agente Usuario, el Agente de Búsqueda por Recuperación y el Agente Información.

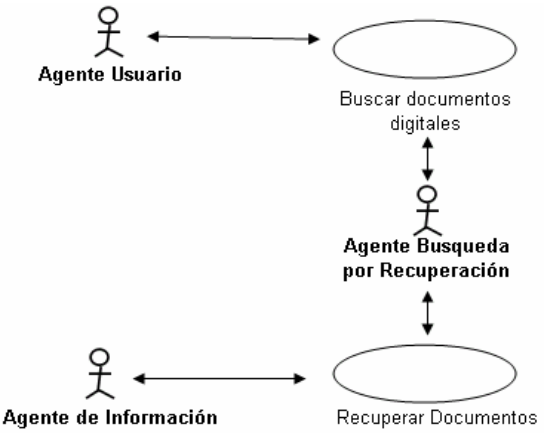

Fig. 4. Caso de uso interno, conversación entre Agentes

Con el fin de complementar la representación de las conversaciones entre los agentes, se hizo de las plantillas textuales, como la descrita en la tabla 4, la cual corresponde a una representación textual de la conversación "*Buscar documentos digitales*" en la que están involucrados los agentes: Usuario y Búsqueda por Recuperación.

Tabla 4. Plantilla textual de conversación: Buscar documentos digitales

Conversación: Buscar Documentos Digitales
Objetivos: Procesar la consulta teniendo en cuenta el modelo de recuperación de información para en etapa posterior realizar la búsqueda en el SRI
Agentes: Agente Usuario, Agente de Búsqueda por Recuperación
Iniciador: Agente Usuario
Servicio: Procesar la consulta realizada por el Agente de usuario
Descripción: El Agente Usuario entrega al Agente de Búsqueda por Recuperación la consulta realizada por el Usuario teniendo en cuenta el modelo de RI
Precondición: El Usuario ha entregado al Agente Usuario la consulta a realizar al Sistema de Recuperación de Información
Postcondición: Se ha procesado la consulta de usuario y cumple con requerimientos impuestos por el modelo de recuperación de información
Condición de Terminación: El Agente de Búsqueda por Recuperación a enviado el resultado de la búsqueda al Agente Usuario para que este se la presente al Usuario

5. FASE DE DISEÑO

Está última fase de la metodología MAS-CommonKADS, se centra en el desarrollo del modelo de diseño. Su objetivo es transformar las especificaciones de los modelos anteriormente relatados para que puedan ser descritos en un lenguaje de programación.

El modelo MAS-CommonKADS tiene tres clases de decisiones de diseño: diseño de la red, diseño de los agentes y diseño de la plataforma.

El diseño de la red consiste en desarrollar el modelo de red, el cual proporciona a los agentes una visión uniforme de la red. Sin embargo S.A.B.I.O, es un sistema cuyos agentes no están en red, por el contrario ellos se encuentran localizados dentro de un solo servidor.

La descripción del diseño de los agentes se realiza siguiendo una plantilla propuesta por MASCommon-KADS. En la tabla 5, se muestra parte de la plantilla del diseño del Agente Indexador, la cual incluye el nombre del agente que se está diseñando, el lenguaje de diseño y los subsistemas (tareas que realiza el agente) encapsulados en el agente. Así mismo los subsistemas, son descritos involucrando el nombre, el tipo y las funcionalidades.

El diseño de la plataforma permite documentar las decisiones de bajo nivel sobre el lenguaje de implementación seleccionado, el software y el hardware empleados y los usuarios finales. Como lenguaje de implementación se escogió JADE (Bellifemine et al, 2004), que es una herramienta *Open-Source* basada en Java que brinda un conjunto de funciones que hacen posible:

- La creación de Agentes
- Programación de las tareas
- Mecanismos de comunicación, específicamente RMI (Remote Method Invocation) (Keogh, 2003).

Tabla 5. Plantilla textual Agente Indexador

Sistema-Agente Indexador
Arquitectura: Los agente se desarrollaron en JADE, sin el uso de ninguna plataforma de construcción de agentes. **Tiene-subsitemas:** Procesar documento, Notificar **Lenguaje-diseño**: JADE **Subsistema:** Procesar documento **Tipo**: Tarea. Subsistema de ejecución de una actividad **Funcionalidades**: Se encarga de realizar el procesamiento de los documentos incluidos en la base documental, este procesamiento se realiza según el modelo vectorial. **Implementa**: Indexar documento

Para la implementación de las interfaces web, se utilizo JSP (Tremblett, 2002) y como lenguaje de representación de los mensajes que intercambian los agentes durante sus conversaciones se usó el lenguaje KQML (Weiss, 1999).

Para detallar un poco más el funcionamiento del sistema en la figura 5, se exhibe la primera interfaz que el "usuario especializado" visualiza al ingresar a S.A.B.I.O., en este caso el sistema actúa de la siguiente manera: el usuario especializado presiona el botón subir archivo, el *Gateway* captura la petición y envía al Agente Usuario Especializado la localización del documento, este agente captura la petición y envía la orden al Agente Indexador para que realice el análisis automático, el Agente Indexador realiza un procesamiento del documento teniendo como base el modelo de recuperación de información vectorial y envía un mensaje al Agente de Base de Datos Documental para que

realice el cálculo de pesos de los términos de los documentos teniendo como base la fórmula especificada en el modelo de recuperación.. Tras haber cumplido por completo el anterior recorrido, el Usuario Especializado, recibe una notificación que confirma el éxito del proceso.

Fig. 5. Interfaz Actor Usuario Especializado

6. CONCLUSIONES

En este artículo, se han explotado las bondades de los sistemas Multi-Agente, tales como: su autonomía y capacidades sociales, para llevar a cabo este proceso de recuperación de información de documentos digitales. A su vez se ha facilitado en gran medida el proceso de modelamiento de S.A.B.I.O., gracias al desarrollo de la metodología MAS-CommonKADS, una de las más completas en lo relacionado con el análisis y diseño de sistemas Multi-Agentes.

La implementación del sistema de almacenamiento y recuperación de información organizada (SABIO) mediante la utilización de la tecnología de agentes de software permite que se utilicen mecanismos de comunicación y colaboración entre entidades autónomas con el fin de satisfacer los requerimientos informativos de los usuarios.

El valor agregado funcional más importante de S.A.B.I.O., es el de ser un sistema de RI que aplica el modelo vectorial, lo suficientemente abierto y flexible para ser implementado, salvo algunos ajustes ajustado (con solo unos pequeños cambios de interfaz) y utilizado en otros centros documentales digitales, que requieran solucionar problemas de búsquedas de información documental..

El reto ahora es complementar la fase de desarrollo de nuestro sistema S.A.B.I.O. anexando un módulo de búsqueda semántico, con el fin de obtener resultados de búsquedas más precisos en unas áreas específicas del saber; este módulo estará compuesto por: bases de conocimiento, representadas por ontologías y módulos de consulta semántica.

REFERENCIAS

Arenas Alvaro E, Barrera Gareth, Pérez Alcázar José De Jesús (2002). Agentes inteligentes para la gestión de memorias organizacionales.

Baeza-Yates. R. y Ribeiro-Neto, B (1999). Modern Information Retrieval. Addison-Wesley. New York.

Bellifemine Fabio, Giovanni Caire, Tiziana Trucco Giovanni Rimassa (2004). JADE Programmer's GUIDE.

FIPA: http://www.fipa.org/specs/fipa00061/, fecha de última visita 21 de mayo, 2005

Hendler, James. Is There an Intelligent Agent in Your Future? http://www.nature.com/nature/webmatters/agents/agents.html, Fecha de última visita: 23-05-2005.

Iglesias, C. (1998). Definición de una metodología para el desarrollo de Sistemas Multi-Agentes. Tesis Doctoral. Departamento de Ingeniería de sistemas telemáticos, Universidad Politécnica de Madrid.

Joyanes Aguilar, Luis, (1998) "Programación orientada a objetos", Madrid Osborne/ McGraw-Hill cop.

Keogh Jim (2003), J2EE Manual de Referencia, Ed. McGraw-Hill, Madrid, pag 803.

Salton, G. (1989). Automatic text processing:the transformation, anlysis and retrieval of information by computer. Reding (MA): Addison-Wesley.

Salton, G., Buckley C. (1988). Term-weighting approaches in automatic text retrieval, Information processing & management, vol.24, num 5. pp 513-523.

Salton, G., Mcgill, J. (1983). Introduction to modern information retrieval. New York: McGraw-Hill.

Schreiber A. Th., Wielinga B. J., Akkermans J.M. and Van de Velde (1994). CommonKADS: A comprehensible methodology for KBS development Deliverable DMI 2ª ADSII/RR/Uva/70/1.1 University of Amsterdam, Netherlands and Free University of Bruselas