Federación distribuida de fuentes de información heterogéneas basada en servicios web y ontologías

David F. Barrero, Diego R. Lopez
RedIRIS
Edificio Bronce
Plaza Manuel Gómez Moreno s/n
28020 Madrid
{david.barrero, diego.lopez}@rediris.es

M. Dolores R-Moreno y Óscar García
Dept. de Automática
Univ. de Alcalá
Ctra. Madrid-Barcelona Km. 31,6
28871 Alcalá de Henares, Madrid
{mdolores, oscar}@aut.uah.es

Resumen

Una de las consecuencias del incremento de la utilización de tecnologías de la información es el incremento del volumen de información generada. Debido a este mayor flujo de información han proliferado un gran número de sistemas que la gestionan, siendo generalmente adecuados sólo al problema concreto al que se dirigen. En determinadas circunstancias es necesario acceder a un conjunto de sistemas de información, lo cual puede originar problemas de interoperatibilidad tanto mayores cuanto más diferente sea la naturaleza de las fuentes accedidas.

En el presente texto se plantea una solución al problema de la integración de información, para lo cual se realiza una traducción semántica de la información utilizando una ontología arbitraria en agentes distribuidos, basados en servicios web, y que pretende ser una solución eminentemente práctica, priorizando su fácil implementación en entornos de producción.

1. Introducción

La gestión de la información es cada vez una tarea más compleja. Por una parte la amplia difusión de las Tecnologías de la Información así como su gran eficiencia que ha inducido un ritmo de generación de información sin precedentes. Por otra parte, el amplio espectro de contextos tecnológicos en los que la información es tratada, cada una con sus particularidades, ha originado la creación de una cantidad considerable de soluciones distintas, generalmente adecuadas para el problema específico que plantean.

A este hecho hay que añadir que la información habitualmente se almacena en formatos poco estructurados, con una escasa metainformación asociada y sin una semántica claramente definida; es evidente que en este contexto la gestión eficiente de la infromación es cada vez más difícil. un ejemplo paradigmático lo podemos encontrar en la Web tradicional, que hace uso intensivo de HTML, un lenguaje de descripción que hace escasa separación entre semántica y representación dificultando la recuperación de información.

A lo largo del artículo presentamos una solución a este problema que se materializa en un software denominado Searchy. Para ello se plantea un entorno de integración de información basado en la idea de federación de servicios, utilizando sistemas de información previamente implementados para dar una interfaz de acceso uniforme. Complementando a esta funcionalidad básica, el sistema se distribuye, de manera que pueda escalar de forma más satisfactoria y pueda adaptarse en entornos B2B o G2G. En la base tecnológica de Searchy se encuentra por una parte los servicios web y por la otra la web semántica.

Searchy ha sido implementada en RedIRIS con notable éxito. A través de Searchy se ha

creado una interfaz de consulta única para las numerosas fuentes de información existentes en RedIRIS. Se ha podido integrar satisfactoriamente el directorio LDAP, listas de correo, repositorios de paquetes software, el contenido del FTP y el contenido del sitio web de RedIRIS. Para ello se ha utilizado el soporte nativo de Searchy así como sistemas de indexación estándar como ffsearch [1] o rpm2html[2].

El artículo se estructura como sigue. Primeramente enmarcaremos el contexto tecnológico que rodea el trabajo para posteriormente describir la motivación del trabajo así como fijar los objetivos concretos que se pretende cumplir. A continuación, la sección 4 presenta la aplicación desde una perspectiva amplia para describir en detalle su arquitectura en la sección 5. La descripción de la aplicación culmina en la sección 6 con una exposición del mecanismo de recuperación e integración de la información. Y finalmente, una reseña de las perspectivas de futuro de la aplicación y unas conclusiones.

2. Marco tecnológico

La integración de información es un campo centro de un interés creciente y que tiende a la incorporación de mecanismos que aporten una cierta inteligencia, cambiando el concepto de información por el de conocimiento [13]. Con el desarrollo de la web semántica [5] la integración de información ha encontrado un nuevo marco de desarrollo impulsado por la estandarización de tecnologías clave en el marco de la Web. El impacto que la web semántica ha tenido en la recuperación de información previsiblemente se verá complementada por otra tecnología emergente: los servicios web.

Particularmente importante en la integración de información está siendo el concepto de ontología [3]. Este término, adoptado de la Filosofía es, según la RAE *"parte de la metafísica que trata del ser en general y de sus propiedades trascendentales"* y que, en su vertiente tecnológica tiene un significado análogo, permite la formalización de marcos conceptuales, o, expresado de otra manera, permite formalizar las propiedades trascendentales de dominios de conocimiento.

La aplicación de las ontologías en la integración de información permite definir modelos compartidos sobre la información a la que se accede y que se pretende integrar, facilitando de esta manera la integración semántica [11]. Con la aparición de la web semántica las ontologías han recobrado gran interés por parte de la comunidad científica, siendo objeto de intenso estudio, y la estandarización de lenguajes de ontologías como RDFS (*RDF Schema*) y OWL (*Web Ontology Language*) suponen un punto de partida crítico para su difusión fuera de los laboratorios.

El trabajo realizado en los últimos años sobre la integración de información [11] se viene centrando en la utilización de ontologías. Mayoritariamente estas aproximaciones son específicas de la aplicación [7] o bien del soporte de la información, siendo difícil su utilización en otros entornos o con distintos soportes de información. También existe interés en la adopción de tecnología de agentes en la integración de información con distintos enfoques. Malucelli [6] creó un agente específico de ontologías para la monitorización de los actos comunicativos. Una solución particularmente interesante es InfoSleuth [8], que proporciona integración basada en ontologías con una notable funcionalidad.

Los servicios web impulsaron la creación de nuevos sistemas distribuidos de integración de información. Knowledge Sifter [9] es un sistema de agentes especializados en distintas tareas de integración de información que utiliza servicios web como interfaz de acceso a ciertas fuentes de información y UDDI (*Universal Description, Discovery, and Integration*) como mecanismo de descubrimiento de los agentes. Otra aproximación basada en servicios similar es utilizado por SODIA [14], que se basa en una arquitectura orientada a servicio (SOA) para realizar la integración de la información.

[1] http://ffsearch.sourceforge.net/
[2] http://rpmfind.net/linux/rpm2html/

3. Motivación

Las aproximaciones realizadas en la recuperación de información son satisfactorias para los objetivos que plantean, si bien no están extentos de ciertas carencias. Por una parte el ámbito al que se circunscriben suele ser reducido, bien por el soporte de información que integran, bien por su especifidad a unas determinadas ontologías y/o soporte de la inforamción, por lo que es difícil su aplicabilidad en problemas similares pero distintos; otros sistemas resultan excesivamente complejos, reduciendo su utilización en ambientes académicos relativamente reducidos.

Nuestra perspectiva intenta aportar una mejora respecto a estos aspectos de la integración de información. Abordamos la problemática desde una perspectiva generalista y dinámica tanto a la integración estructural como semántica. Este hecho implica no adoptar una ontología determinada y permitir una fácil extensión a distintos soportes de información; así mismo se considera crítico ofrecer un compromiso entre funcionalidad y sencillez de uso, de manera que se proporcione un máximo de funcionalidad con la menor complejidad de implementación posible.

En la solución propuesta no sólo se considera el problema intrínseco de la integración descrito, sino que consideraciones contextuales tienen un peso específico fundamental. La integración de información es de por si un problema de primer orden que puede aumentar en órdenes de magnitud si consideramos el hecho de que no se tenga un pleno control sobre las fuentes de información, situación arquetípica de entornos con interacciones horizontales como en B2B. Ofrecer una solución adecuada a esta problemática implica considerar tanto condicionantes técnicos como de otras índoles.

Las consideraciones técnicas son las propias de un entorno B2B, con los problemas de interoperatibilidad y portabilidad, que puede solucionarse de forma satisfactoria con una adecuada elección de la tecnología. Analizando la problemática desde una perspectiva más general podemos plantear que en un entorno multiorganizacional como el presentado plantear soluciones centralizadas y de cierta complejidad puede generar resistencias de diversa índole que imposibiliten su satisfactoria implementación.

En este contexto puede resultar más adecuado una aproximación distribuida, lo que va a permitir que:

- El mantenimiento de la aplicación requiera de una mínima autoridad central.
- Todos los entes implicados participan en el éxito de la aplicación desde posicones de igualdad (Colaboración).
- Se aproveche los sistemas de información en funcionamiento dentro de una organización, minimizando las redundancias y maximizando su eficiencia (Federación).
- Un bajo grado de acoplamiento con el sistema de integración (No intrusismo).

Planteamos dos objetivos fundamentales en la concepción de Searchy: Simplificar el sistema cuanto sea posible, con el objeto de facilitar al máximo su despliegue por parte de personal no especializado, esto implica renunciar a ciertas funcionalidades que pueden resultar interesantes, pero que perjudican la sencillez de mantenimiento del sistema. Por otra parte, se pretende proporcionar una importante generalidad respecto a las fuentes de datos que puedan ser accedidas, permitiendo una fácil extensión al sistema.

4. Descripción de Searchy

El proyecto Searchy [10] nació con la vocación de hacer una aportación en la solución de parte del problema descrito, integrando información con un objetivo específico: localizar y describir recursos heterogéneos ubicados en distintos dominios administrativos [1]. Para ello Searchy estaba ligado a la utilización de una ontología concreta, Dublin Core, que servía como modelo de datos abstracto para la descripción de los recursos. El éxito de la aproximación adoptada y la experiencia acumulada sugirió la conveniencia de adoptar un espectro más amplio de aplicaciones, como en la

Gestión del Conocimiento dentro de la Administración Pública [2].

Una evolución posterior de Searchy permitió su utilización junto a ontologías arbitrarias, transformándolo de metabuscador distribuido a un sistema completo de integración de sistemas de información generalista, abarcando tanto la integración estructural como semántica de fuentes arbitrarias de información que incluye, aunque no se limita, a bases de datos, directorios, índices y buscadores web.

Funcionalmente Searchy recibe una consulta formulada abstracta independiente del soporte final, traduce la consulta a un formato comprensible por un o unos sistemas de información locales, la envía, procesa la respuesta devuelta integrándolas y mapeándolas a un formato común.

Searchy está basado en el concepto de agente [12], que le otorga un elevado grado de distribución así como una arquitectura flexible. La implementación se ha realizado siguiendo estándares del W3C, usando intensivamente dos tecnologías: Los servicios web para el acceso a los servicios y la web semántica para el acceso a la información. Los agentes exponen sus servicios como servicios web accesibles a través de SOAP y la información integrada se intercambia en formato RDF (*Resource Description Framework*).

Cabe destacar que Searchy es una capa de middleware y que, por lo tanto, no está orientado a ser utilizado por usuarios finales, sino por otras aplicaciones. La publicación del servicio en forma de servicio web permite que Searchy sea utilizado dentro de un espectro amplio de entornos, desde aplicaciones web hasta aplicaciones pesadas; los clientes Searchy pueden funcionar como una mera interfaz gráfica al núcleo de Searchy recogiendo la consulta del mismo y visualizando los resultados, pero también puede ser utilizado por aplicaciones como fuente de datos con las que realizar tareas no directamente relacionadas con el usuario final. Existe un amplio espectro de aplicaciones potenciales.

5. Arquitectura

El elemento básico en la arquitectura de Searchy es el agente, que es la mínima unidad con funcionalidad completa. Los agentes no están especializados, todos realizan las mismas tareas con una salvedad: cada agente puede acceder a distintos sistemas de información. Este hecho permite una notable flexibilidad operativa, de manera que pueden implementarse distintas políticas de alto nivel diseñando un despliegue de agentes adecuado.

Tal y como se puede visualizar en la figura 1, cada agente se divide en tres elementos arquitecturales y de cuyas características se derivan algunas propiedades fundamentales de los agentes. A continuación se describen estos tres elementos.

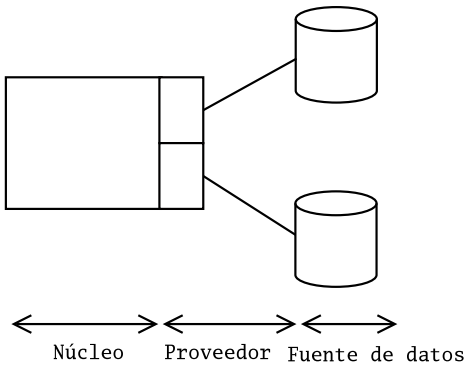

Figura 1: Estructura de un agente Searchy

- Núcleo. Contiene los elementos comunes a todos los agentes, se encarga de tareas tales como el procesamiento del mensaje SOAP, manejo de la configuración del agente, proporcionar servicios básicos tanto al cliente del agente como al proveedor y, en definitiva, de todo aquello que no sea específico de las fuentes de datos.
- Proveedor. Sirve de interface entre el núcleo del agente y la fuente de datos implementando el algoritmo de acceso al formato local de información. Cada tipo de fuente de datos requiere un tipo de

proveedor específico. Facilita la extensividad del sistema, de manera que se puedan implementar el soporte a nuevas fuentes de datos con relativa sencillez.

- Fuente de datos. Es el soporte final de la información que se integra. Como fuente de datos se puede utilizar cualquier recurso que almacene información digital, aunque dado el enfoque del sistema la fuente de datos normalmente será un sistema de información como una base de datos o un índice.

El soporte de nuevas fuentes de datos se realiza por medio del desarrollo de proveedores específicos, pudiendo el proveedor implementar cualquier algoritmo para acceder a los datos. Esta característica permite la integración de datos que existen dentor de un soporte de información, pero también permite generar la información dinámicamente, por ejemplo, accediendo a sensores o incorporando un algoritmo de recuperación de información.

El estado de desarrollo actual de Searchy abarca cuatro tipos de proveedores: SQL, LDAP, Google y Harvest. Por medio de los proveedores SQL y LDAP se puede acceder a datos estructurados de bases de datos relacionales y directorios que suelen conforman el backend de un sistema de información. El proveedor Google utiliza el API SOAP que ofrece Google para que las aplicaciones puedan interaccionar directamente con su motor de búsqueda, así se pueden realizar búsquedas que abarquen toda internet o un determinado sitio web. Por último, a través del indexador Harvest se pueden localizar recursos como las páginas web o documentos LaTeXde una intranet.

Una característica fundamental de los agentes es que tienen capacidad de intercomunicarse, cooperando para atender una petición. Un despliegue de Searchy típico consistirá en distintos agentes, cada uno accediendo a una o varias fuentes de información.

El criterio para asignar responsabilidades a los agentes puede ser de muy diversa índole, y generalmente corresponderá a consideraciones no estrictamente técnicas; típicamente este criterio responderá a razones administrativas, de manera que cada agente proporcione integración a una determinadad unidad organizativa, una universidad o un departamento, por ejemplo.

La figura 2 muestra un ejemplo de cómo puede realizarse un despliegue de agentes Searchy dentro de una universidad; cada departamento tendría una completa autonomía a la hora de adoptar decisiones tecnológicas, y, una vez realizada esta, debería situar un agente que garantize la accesibilidad exterior de los datos. Los distintos agentes integran los datos de todos los sistemas dando una visión homogenea de los mismos; no obstante, es posible diferenciar la procedencia de la información añadiendo campos al modelo abstracto.

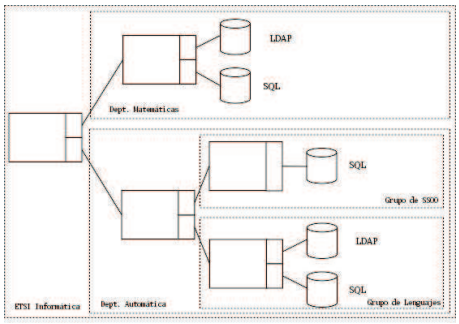

Figura 2: Ejemplo de despliegue de Searchy

6. Recuperación de información

Un aspecto crítico es el formato de representación de la información. Debe ser un formato independiente de la plataforma y, dada la generalidad de usos que puede recibir se requiere que el formato sea semánticamente flexible y autocontenido. Estas condiciones son imprescindibles para un entorno heterogéneo como el planteado y una flexibilidad operativa que permita adaptarse a distintos tipos de información de manera sencilla.

Una tecnología que satisface estas necesidades se puede encontrar en la web semán-

tica. Searchy utiliza RDF para representar la información junto con una codificación en XML que nos facilita la interoperatibilidad. Estas tecnologías permiten el intercambio estructural de la información, pero no el intercambio de su semántica, para ello se requiere una formalización de la ontología proporcionada por RDFS y OWL. Searchy no requiere un conocimiento de la ontología utilizada, le basta con poder realizar mapeos entre las ontologías locales manejadas por cada fuente de información y la ontología compartida, que típicamente será consensuada entre las partes implicadas en un despliegue.

El mapeo entre ontologías es un aspecto relevante que en sí mismo es objeto de investigación. El perfil eminentemente práctico de Searchy ha conducido a adoptar un enfoque conservador a este problema, ofreciendo una interfaz al administrador del agente para que defina los mapeos manualmente. Para ello se proporciona un mecanismo sencillo de substitución de cadenas, de manera que una parte considerable de la complejidad queda encapsulada y con un mínimo esfuerzo pueden establecerse las reglas de mapeo.

7. Trabajo futuro

Un problema intrínseco a la naturaleza distribuida del trabajo descrito es la fuerte dependencia del tiempo de respuesta con el transporte, lo que limita la escalabilidad del sistema dado el enfoque estático con el que se ha realizado la implementación. La utilización de servicios web permite adoptar una variedad considerable de mecanismos de transporte adecuados a distintos entornos. Unido al transporte está la problemática del descubrimiento de agentes, problema análogo al descubrimiento de servicios web, y por lo tanto se plantea la conveniencia de utilizar directorios UDDI así como otros mecanismos más exóticos, como es la utilización de SOAP con multicast por IP o protocolos peer-to-peer como JXTA. La incorporación de Searchy en dispositivos móviles como teléfonos o PDA también se presenta como una excelente perspectiva de futuro.

Por otra parte, la creación de una red de agentes extensa puede implicar manejo volúmenes de información muy considerables, hecho que puede ser problemático en un entorno de búsqueda de información, por lo tanto se hace imprescindible acompañar la escalabilidad física del transporte con mecanismos que permitan precisar la búsqueda cuanto sea posible. Con este fin se plantea el desarrollo de clientes inteligentes basados en ontologías.

Un aspecto a considerar es que la información puede ser de naturaleza sensible por lo tanto puede resultar necesario habilitar mecanismos de autentificación así como control de derechos de acceso, para lo cual se pretende incorporar SAML (Security Assertion Markup Language).

Para que Searchy sea útil en entornos de producción reales es necesario que sea capaz de acceder a sistemas de información diversos, por lo tanto el desarrollo de nuevos proveedores que añadan soporte a dichos sistemas es un aspecto crítico.

También se estima necesario mejorar la comprensión de los entornos en los cuales Searchy puede ser potencialmente utilizado. La flexibilidad de su enfoque permite su utilización con una cantidad considerable de distintas políticas y objetivos distintos, así se plantea, por ejemplo, su utilización como localizador de presencia o de objetos distribuidos. Profundizar en esta cuestión se desvela como un aspecto clave para llevar a Searchy al máximo de su potencial.

8. Conclusiones

A lo largo del presente artículo se ha presentado brevemente la problemática de la integración de información y se introduce una solución parcial con un perfil marcadamente práctico denominado Searchy. Esta solución, basada en la federación de servicios, está orientada a entornos distribuidos complejos, en los que sea necesario respetar la independencia tecnológica de cada una de las partes implicadas.

El coste de adopción de Searchy es muy limitado, está publicado bajo licencia GPL y no requiere una reconversión de la infraestructura

tecnológica, sino que se adapta a los medios ya existentes.

La flexibilidad y generalidad con el que se ha concebido nos permite afirmar que en un futuro se encontrará aplicaciones dificilmente previsibles cuando se ideó Searchy.

9. Agradecimientos

El proyecto Searchy es deudor de la colaboración de muchas personas; cabe destacar y agradecer especialmente las aportaciones realizadas por José Manuel Macias, Ajay Arjandas Daryanani y Javier Masa Marín, todos miembros de RedIRIS. Este trabajo ha sido financiado pro el proyecto de la Universidad de Alcalá PI2005/084 en colaboración con RedIRIS a través del programa PTYOC [4].

Referencias

[1] Barrero D. F., López R. D. y García Población, O. *Distributed Metainformation Searching: an Approach to Information Retrieval in the Age of the Semantic Web.* En VIIth Trans-European Research and Education Networking Association Networking Conference. Rodas, Grecia, junio 2004.

[2] Barrero D. F., Criado, J. I. *Integrando la Información de las Administraciones Públicas en la Gestión del Conocimiento. Una Solución desde la Web Semántica y los Servicios Web.* En VIII TECNIMAP. Murcia, España, septiembre-octubre 2004.

[3] Gómez Pérez, A., Fernández-López, M. y Corcho, O. *Ontological Engineering.* Springer-Verlag, 2003.

[4] RedIRIS *Colaboración de RedIRIS en la realización de trabajos académicos*, RedIRIS, 2005. http://www.rediris.es/app/PTYOC.

[5] Hendler, J., Berners-Lee, T., Lassila, O. *The semantic web.* Scientific American, 284(5):28-31, mayo 2001.

[6] Malucelli, A., Oliveira, E. *Ontology-services agent to help in the structural and semantic heterogeneity.* En 5th IFIP Working Conference on Virtual Enterprises. Toulouse, agosto 2004.

[7] Michalowski, M. et alia *Retrieving and Semantically Integrating Heterogeneous Data from the Web.* IEEE Intelligent Systems, Vol. 19, No. 3, pp. 72-79, mayo - junio 2004.

[8] Nodine M., Fowler J. y Perry B. *Active Information Gathering in InfoSleuth.* International Journal of Cooperative Information Systems. Vol 9, Num. 1-2, páginas 3-28. 200.

[9] Kerschberg L. et alia *Knowledge Sifter: Ontology-Driven Search over Heterogeneous Databases.* En 16th International Conference on Scientific and Stadistical Database Management 04. Santorini, Grecia, 2004.

[10] Sitio web del proyecto Searchy. http://jsearchy.sourceforge.net. Fecha de acceso: 6-4-2005

[11] Wache H., et alia *Ontology-Based Integration of Information - A Survey of Existing Approaches.* En Proceedings de IJCAI-01, Seattle, WA, 2001.

[12] Wan, F. y Singh, M. P. *Commitments and Causality for Multiagente Desing.* En 2nd International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS), Melbourne, ACM Pres, julio 2003.

[13] Zaïane, O. R. *From Resource Discovery to Knowledge Discovery on the Internet.* Simon Fraser University, no. TR 1998-13. 1998.

[14] Zhu F., Kotsiopoulos I., Bennett K., Russell M., Budgen D., Brereton P., Keane J., Rigby M., Xu J. *Dynamic Data Integration Using Web Services.* En 2nd IEEE International Conference on Web Services (ICWS'04). San Diego, California, EEUU, 2004.