Aplicaciones de Búsquedas por Similtud en un Portal GRID Orientado a Aspectos*

Osiris Sofía, Sandra Casas
Universidad Nacional de la Patagonia Austral
Lisandro de la Torre 1070
CP 9400. Río Gallegos, Santa Cruz, Argentina
Tel/Fax: +54-2966-442313/17
{osofia;lis}@unpa.edu.ar

y

Roberto Uribe Paredes
Departamento de Ingeniería en Computación
Universidad de Magallanes
Punta Arenas, Chile
ruribe@ona.fi.umag.cl

Resumen

En el marco del Proyecto de Investigación *Paralelización de Estructuras de Datos y Algoritmos para la Recuperación de Información*, de la Universidad Nacional de la Patagonia Austral se desarrolla la línea de investigación que da continuidad al trabajo previo de dos grupos de investigación de esa Universidad en las áreas de paralelismo y programación orientada a aspectos, integrados en esta oportunidad en la búsqueda de alternativas para la implementación del paradigma de orientación a aspectos en entornos paralelos, particularmente en entornos GRID, a través del desarrollo de un *Portal* GRID, utilizando técnicas y herramientas de la programación orientada a aspectos, y del área de búsquedas por similitud investigadas por un grupo de la Universidad de Magallanes, Chile, con el fin de desarrollar aplicaciones arqueológicas para el Portal GRID.

Palabras claves: Computación GRID, Programación Orientada a Aspectos, Portales GRID, Paralelismo, Búsquedas por Similitud.

1. CONTEXTO

En los últimos años, de manera separada han surgido y crecido dos líneas de investigación en la Unidad Académica Río Gallegos de la UNPA. A partir del año 1999 se ha trabajado sobre problemas de la distribución y paralelización de bases de datos y desde el año 2005 un grupo diferente de investigadores ha estudiado el paradigma orientado a aspectos en la búsqueda de estrategias de resolución de conflictos. Por otra parte desde el comienzo de la investigación en paralelismo se ha trabajado en colaboración con la Universidad de Magallanes, Chile, sede Punta Arenas. Actualmente ese grupo desarrolla sus investigaciones en temas relacionados con las búsquedas por similitud. Los tres grupos se encuentran en un estado de consolidación suficiente para confluir sus esfuerzos en un proyecto que unifique sus líneas de investigación.

2. INTRODUCCIÓN

La *Computación Grid* [10] se ha establecido como un nuevo paradigma para la computación científica de gran escala (o redes de investigación): la aplicación de recursos computacionales coordinados, interconectados vía una red pública de alta velocidad, para solucionar problemas en áreas científicas tales como: astrofísica, química, física, geofísica, meteorología y climatología, neurobiología, biología molecular, etc. Las aplicaciones Grid son sistemas computacionales distribuidos que proveen mecanismos para compartir de manera controlada recursos computacionales. La Computación Grid requiere componentes middleware genéricos, que oculten a las aplicaciones específicas los detalles de acceso

*Este trabajo fue financiado por la Universidad Nacional de la Patagonia Austral, Santa Cruz, Argentina, proyecto 'Paralelización de Estructuras de Datos y Algoritmos para la Recuperación de Información'

y usar configuraciones de recursos heterogéneos, procesadores, almacenamiento y conexiones de redes. Estos garantizan la interoperabilidad de los recursos a través del uso de protocolos estándares. El término *Tecnología Grid* [14] usualmente se refiere a éste tipo de middleware.

Uno de los enfoques más utilizados para proporcionar acceso a la Computación GRID son los Portales GRID. Los Portales GRID [7] son herramientas muy eficaces que proporcionan a los usuarios de Computación GRID interfaces simples e intuitivas para el acceso a la información y recursos GRID [10]. La construcción de un Portal GRID debe cumplir con todos los requerimientos de servicios y recursos, para lo cual se han desarrollado APIs, Toolkits, Frameworks específicos, tales como GridPort [12], GPDK [11] y P-GRADE [3].

La Programación Orientada a Aspectos [8] (POA) es un relativamente nuevo paradigma para el desarrollo de software que proporciona abstracciones para la implementación de los crosscutting concerns, de manera separada y aislada a los componentes de funcionalidad básica. La POA además proporciona mecanismos para la composición de las diferentes unidades. En otras palabras, la Orientación a Aspectos, es una técnica que permite aplicar el principio de Separación de Concerns [6] y de esta forma, obtener los beneficios enunciados por dicho principio. La unidad de implementación que representa a la funcionalidad transversal se denomina aspecto, dando origen al nombre del paradigma. De esta forma, se suele referir casi sin distinción para indicar el mismo concepto a los términos aspecto, funcionalidad transversal y/o "crosscutting concern".

Ciertas funcionalidades y requerimientos han sido identificados como "clásicos" crosscutting concerns. Entre estos suelen identificarse: coordinación, distribución, sincronización, concurrencia, balance de carga, seguridad, logging y autentificación. En el desarrollo e implementación de portales GRID, todos o algunos de estos concerns estarán presentes, por lo que se puede suponer a priori, que al desarrollar una aplicación GRID pueden ser implementados bajo el enfoque POA.

2.1. Portales GRID

Un portal GRID es en esencia una aplicación WEB, por lo cual tiene requerimientos (o características) similares a los portales orientados al consumo o usuario (Yahoo, CNN, IBM intranet). Estos servicios suelen incluir soporte para el contexto (login, customización, personalización, etc.); soporte para interfaces de usuario basadas en navegadores; páginas dinámicas disponibles a usuarios anónimos o autenticados. En particular los portales de e-Science deben además soportar cuestiones relacionadas con la integración de aplicaciones de dominio específicas basadas en GRID. Aquí surge la principal diferencia, un portal GRID debe manejar computación que se ejecute por días o semanas sobre cientos de nodos para procesar terabytes de datos científicos. Específicamente estos portales son requeridos para manejar credenciales, lanzar trabajos, manejar ficheros y ocultar la complejidad de la GRID como los trabajos batch distribuidos. En la Figura 1 se comparan ambos tipos de portales [17].

	Grid	Webs	Comentarios
Principales Usos / Usuarios	eScience, eEngineering	Comunicación científica (inicialmente), eCommerce, eContent (multimedia)	Existen algunos solapamientos y habrán más en el futuro.
Principales funciones	Computación de alta performance, compartición de recursos computacionales	Información, comunicación, transacciones	
Aplicaciones	Problemas que requieren computo intensivo en ciencias e ingeniería	Servicios I&C, educación & entrenamiento, eBusiness, eCommerce (B2B, B2C, B2A, etc.), etc.	Las Webs son interfaces principales para acceso a las aplicaciones
Volúmenes de Datos	XXL (y mayores)	S - XL	Grid futuras pueden también trabajar sobre volúmenes más pequeños.
Recursos	Almacenamiento, ancho de banda, tiempo de procesador, ficheros, etc.	Contenido digital y relacionado a servicios.	Contenedores, transportadores & Procesadores vs. Contenido y aplicaciones
Usuarios	Grupos de usuarios especiales (científicos, ingenieros)	Publico en general, negocios, administradores, etc.	Estos son solo algunos de los grupos objetivo.
Standares	Faltan stándares para middleware	Existen algunos stándares y recomendaciones	Las comunidades Grid y Web están todavía lejanamente separadas.

Figura 1: *Comparación de Portales Grids y Portales Webs Orientados al consumo/usuario.*

A continuación se detallan los servicios básicos que un Portal GRID debe ofrecer y el modelo arquitectónico de base empleado en el diseño e implementación de los mismos.

2.2. Servicios soportados por un Portal GRID

Los servicios que un Portal GRID típicamente incluye son:
- *Seguridad*: los usuarios se loguean en un portal usando un navegador WEB y se autentifican mediante un user-id y password. El portal GRID

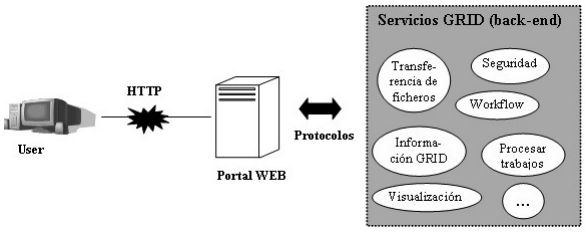

Figura 2: *Arquitectura Base de Portales Grid.*

mapea el user-id a credenciales GRID.
- *Gestión de Datos*: provee acceso a ficheros, colecciones y metadatos locales y remotos, soporta transferencia de ficheros;
- *Job submission*: se refiere a la habilidad de que los procesadores conectados a la GRID ejecuten un trabajo (secuencial o paralelo) y puedan monitorear su estado. Este es un servicio clásico soportado por el portal;
- *Servicios de Información*: el acceso a directorios y estado de herramientas es un rol esencial del Portal;
- *Interfaces de Aplicación*: permite ocultar convenientemente los detalles GRID detrás de una interfaz de aplicación;
- *Colaboración*: los portales sirven como entradas a organizaciones virtuales para compartir recursos;
- *Workflow*: presenta los usuarios con sus tareas y asume la responsabilidad de integrar estas tareas en secuencias;
- *Visualización*: provee herramientas que ofrecen a los usuarios acceso a los datos, renderización y visualización de recursos. Puede proveer algún nivel de vista de los datos o puede ser usada para ofrecer herramientas más avanzadas.

2.3. Arquitectura base de portales GRID

Un Portal GRID se puede ver como una interface WEB a un sistema distribuido. La arquitectura básica de un portal responde a un esquema de arquitectura de tres capas: (1) la capa cliente, que se ejecuta mediante un navegador WEB; (2) la capa servidor que cumple la función de representar la lógica del negocio y (3) la capa de recursos y servicios GRID. Los clientes y el servidor típicamente se comunican vía HTTP permitiendo que cualquier navegador WEB sea usado. La capa servidor simplemente accede a ficheros locales para servir páginas pero también puede dinámicamente generar páginas web mediante la ejecución de scripts CGI y/o mediante interacción directa o indirecta con los recursos back-end. La interacción con la tercera capa puede lograrse en algún protocolo o de manera apropiada. Usando esta arquitectura general, los portales pueden ser construidos para que soporten aplicaciones de una amplia variedad. Para hacerlo efectivamente, sin embargo, se requiere un conjunto de herramientas de construcción de portales que puedan ser personalizados para cada área de aplicación. En la Figura 2 se presenta gráficamente una arquitectura de tres capas aplicada al acceso y utilización de recursos y servicios GRID.

2.4. Búsquedas por similitud en espacios métricos

Uno de los problemas de gran interés en ciencias de la computación es el de "búsqueda por similitud", es decir, encontrar los elementos de un conjunto más similares a una muestra. Esta búsqueda es necesaria en múltiples aplicaciones, como ser en reconocimiento de voz e imagen, compresión de video, genética, minería de datos, recuperación de información, etc. En casi todas las aplicaciones la evaluación de la similaridad entre dos elementos es cara, por lo que usualmente se trata como medida del costo de la búsqueda la cantidad de similaridades que se evalúan. Interesa el caso donde la similaridad describe un espacio métrico, es decir, está modelada por una función de distancia que respeta la desigualdad triangular. En este caso, el problema más común y difícil es en aquellos espacios de "alta dimensión" donde el histograma de distancias es concentrado, es decir, todos los objetos están más o menos a la misma distancia unos de otros. El aumento de tamaño de las bases de datos y la aparición de nuevos tipos de datos sobre los cuales no interesa realizar búsquedas exactas, crean la necesidad de plantear nuevas estructuras para búsqueda por similaridad o búsqueda aproximada. Asimismo, se necesita que dichas estructuras sean dinámicas, es decir, que permitan agregar o eliminar elementos sin necesidad de crearlas nuevamente, así como también que sean óptimas en la administración de memoria secundaria. La necesidad de procesar grandes volúmenes de datos obligan a aumentar la capacidad de procesamiento y con ello la paralelización de los algoritmos y la distribución de las bases de datos. El objetivo de los algoritmos de búsqueda es minimizar la cantidad de evaluaciones de distancia realizadas para resolver la consulta. Los métodos para buscar en espacios métricos se basan principalmente en dividir el espacio empleando la distancia a uno o más objetos seleccionados. El no trabajar con las características particulares de cada aplicación tiene la ventaja de ser más general, pues los algoritmos funcionan con cualquier tipo de objeto [4]. Existen distintas estructuras para buscar en espacios métricos, las cuales pueden ocupar funciones discretas o continuas de distancia. Algunos son BKTree [2], GNAT [1], VpTree [23], MTree [5],

SAT [13], EGNAT [22]. Algunas de las estructuras anteriores basan la búsqueda en pivotes y otras en clustering. En el primer caso se seleccionan pivotes del conjunto de datos y se precalculan las distancias entre los elementos y los pivotes. Cuando se realiza una consulta, se calcula la distancia de la consulta a los pivotes y se usa la desigualdad triangular para descartar candidatos.

3. LÍNEA DE INVESTIGACIÓN Y DESARROLLO

El objetivo del trabajo es demostrar empíricamente que la orientación a aspectos es una técnica de desarrollo de software más conveniente para la construcción de Portales GRID que las técnicas convencionales de desarrollo de software (orientadas a objetos y/o componentes).
La hipótesis principal consiste en que el diseño e implementación de un Portal GRID empleando la orientación a aspectos genera una aplicación más reutilizable, mantenible, escalable, evolucionable y traceable.
El método de investigación es eminentemente empírico, para lo cual se seleccionará un caso de estudio particular y real. Para este caso concreto se desarrollará una aplicación consistente en un simulador arqueológico que implementará además búsquedas por similitud.
Hasta el momento se han realizado las etapas de *Estudio del Estado del Arte*, *Identificación de requerimientos* y *Diseño Arquitectónico*. Las etapas pendientes se describen a continuación.
Implementación: En esta etapa se debe en principio seleccionar las herramientas de programación más adecuadas, garantizado la compatibilidad entre las mismas. El lenguaje de componentes para la funcionalidad base y un lenguaje orientado a aspectos, además del servidor web y herramientas para servicios específicos del Portal GRID.
Pruebas: En principio se establecen pruebas funcionales que garanticen el correcto funcionamiento del Portal GRID, pero además se establecerán pruebas de performance y rendimiento.
Comparación: En esta etapa se pretende realizar diversos estudios comparativos con portales GRID desarrollados bajo enfoques diferentes.

4. RESULTADOS OBTENIDOS / ESPERADOS

Durante el primer año del proyecto se han realizado esfuerzos tendientes al cabal entendimiento de las distintas tecnologías que se pretende integrar. De este manera distintos subgrupos han abordado la investigación y publicación de temas tales como GRID [18, 19, 21], Portales GRID [15], simuladores grid, desarrollo de aplicaciones paralelas utilizando la programación orientada a aspectos [9], y paralelización de estructuras de búsqueda por similitud [16].
Durante esta segunda etapa se espera integrar las distintas tecnologías, de manera de implementar aplicaciones reales que realicen búsquedas por similitud en el entorno de un Portal GRID desarrollado bajo el modelo de programación orientada a aspectos.

5. FORMACION DE RECURSOS HUMANOS

La integrante del grupo de investigación de la Universidad Nacional de la Patagonia Austral Natalia Bibiana Trejo se encuentra desarrollando a través de una beca obtenida de la Fundación Carolina y la propia Universidad, el doctorado en Informática otorgado por la Universidad Complutense de Madrid, España. Para la obtención de dicho título es requisito obtener previamente el de Master en Investigación en Informática, lo que ha realizado en el transcurso del año 2008 con la tesis denominada 'Aplicación de Multicast IPv6 Seguro a Servicios de Información en Entornos Grid', bajo la dirección del Dr. Juan Carlos Fabero Jiménez [20]. Actualmente continúa sus estudios para obtener el título de doctorado.
Durante el año 2008 se ha postulado un alumno como becario del proyecto, con el objetivo de desarrollar una aplicación que valide empíricamente el modelo propuesto para el portal GRID. El alumno Franco Herrera desarrollará una aplicación consistente en un simulador arqueológico.
Se ha promovido la participación de los auxiliares de docencia a través de la presentación de los trabajos aprobados en los distintos congresos donde se realizaron aportes, de manera de fomentar su interés no sólo en la etapa de investigación sino también de la transferencia de los resultados y la comunicación con otros grupos del país y del exterior.
Actualmente dos alumnos de la carrera de Licenciatura en Sistemas participan como integrantes en el proyecto de investigación.
Por último durante el 2008 se han realizado pasantías entre integrantes de los grupos de la Universidad Nacional de la Patagonia Austral y la Universidad de Magallanes para intercambiar habilidades en la aplicación de las áreas de especialización de cada grupo.

Referencias

[1] Sergei Brin. Near neighbor search in large metric spaces. In *the 21st VLDB Conference*, pages 574–584. Morgan Kaufmann Publishers, 1995.

[2] W. Burkhard and R. Keller. Some approaches to best-match file searching. *Communication of ACM*, 16(4):230–236, 1973.

[3] Nemeth C., Dozsa G., Lovas R., and Kacsuk P. The p-grade grid portal. *LNCS, v. 2044*, pages 10–19, 2004.

[4] Edgar Chávez, Gonzalo Navarro, Ricardo Baeza-Yates, and José L. Marroquín. Searching in metric spaces. In *ACM Computing Surveys*, pages 33(3):273–321, September 2001.

[5] P. Ciaccia, M. Patella, and P. Zezula. M-tree : An efficient access method for similarity search in metric spaces. In *the 23st International Conference on VLDB*, pages 426–435, 1997.

[6] Dijkstra E. *A Discipline of Programming*. Prentice-Hall, 1976.

[7] Furmanski W. Fox G. *High performance commodity computing*, chapter 10. Morgan Kaufman Publishers, 1998.

[8] Kiczales G., Lamping L., Mendhekar A., Maeda C., Lopes C., Loingtier J., and Irwin J. Aspect-oriented programming. In *In Proceedings ECOOP 97*, Finland, 1997.

[9] Esteban Gesto, Sandra Casas, and Osiris Sofia. Implementación de aplicaciones paralelas bsp usando aspectos. In *XXIII Brazilian Symposiums on DataBases (SBBD) and XXII Brazilian Symposiums Software Engineering (SBES) II Latin American Workshop on Aspect-Oriented Software Development - LA-WASP 2008*, pages 96–97, october 2008.

[10] Foster I. and Kesselman C. *The Grid: Blueprint for a New Computing Infrastructure*. Morgan Kaufman Publishers, 1998.

[11] Novotny J. Grid computing environments special issue of concurrency and computation practice and experience. *The Grid Portal Development Kit*, pages 1129–1144, 2002.

[12] Thomas M., Mock S., and Boisseau J. Development of web toolkits for computational science portals: The npaci hotpage. In *9th IEEE International Symposium on High Performance Distributed Computing*, pages 308–309, 2000.

[13] Gonzalo Navarro. Searching in metric spaces by spatial approximation. *The Very Large Databases Journal (VLDBJ)*, 11(1):28–46, 2002.

[14] Proyecto Globus. http://www.globus.org/.

[15] Albert Osiris Sofia and Sandra Isabel Casas. Survey de tecnologías grid. *Revista Ciencia y Técnica Administrativa*, 07(04), 2008. http://www.cyta.com.ar/ta0704/v7n4a1.htm.

[16] Roberto Solar-Gallardo, Roberto Uribe-Paredes, Esteban Gesto, and Osiris Sofia. Implementación de un digesto digital paralelo para búsquedas por similtud sobre documentos. In *XIV Congreso Argentino de Ciencias de la Computación - CACIC 2008*, october 2008.

[17] Hans-Georg Stork. Webs, grids and knowledge spaces, - programmes, projects and prospects. In *Journal of Universal Computer Science*, volume 8, pages 848–868, 2002.

[18] Natalia Trejo and Juan C. Fabero. Aplicación de multicast ipv6 a servicios de información en entornos grid. In *XXXIV Conferencia Latinoamericana de Informática - CLEI 2008*, pages 183–192, september 2008.

[19] Natalia Trejo and Juan C. Fabero. Grid resource discovery using signed ipv6 multicast. In *Jornadas Chilenas de Computación 2008. XII Workshop de Sistemas Distribuidos y Paralelismo. JCC 2008*, pages 20–21, november 2008.

[20] Natalia Bibiana Trejo. Aplicación de multicast ipv6 seguro a servicios de información en entornos grid. Tesis de máster en investigación en informática, Facultad de Informática, Universidad Complutense de Madrid, 2008.

[21] Natalia Bibiana Trejo and Juan Carlos Fabero Jiménez. Descubrimiento de recursos grid utilizando multicast ipv6 seguro. *Revista de Ingeniería Informática, JCC 2008. Selección de los mejores artículos del Workshop de Sistemas Distribuidos y Paralelismo*, 01(01), 2008. http://revista.inf.udp.cl/index.php.

[22] Roberto Uribe-Paredes. Manipulación de estructuras métricas en memoria secundaria. Master's thesis, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Santiago, Chile, Abril 2005.

[23] P. Yianilos. Data structures and algoritms for nearest neighbor search in general metric spaces. In *4th ACM-SIAM Symposium on Discrete Algorithms (SODA'93)*, pages 311–321, 1993.