SERIE: SISTEMAS DE ALMACENAMIENTO DISTRIBUIDO (III)

Mapas: representación de los SAD de acuerdo con la naturaleza de su información[1]

Ricardo Marcelín Jiménez

Agosto de 2011

El tipo de información que una organización genera define, en gran medida, su carácter. Dicha información está sometida a determinadas condiciones culturales y supuestos teóricos de acuerdo con el lugar, el momento y la perspectiva desde la que se genera. En suma, un grupo de trabajo comparte documentos de interés común con un léxico especializado y, por lo tanto, relacionados implícitamente. No obstante, no todos esos archivos comparten las mismas situaciones: hay documentos que se usan y cambian constantemente, y otros que tardan en volver a ser consultados. En la planeación de un Sistema de Almacenamiento Distribuido (SAD), hoy en día resulta básico considerar el tipo de información que cada organización maneja y el uso que le brinda; es decir, no es igual la información empleada por una organización de seguridad nacional que otra a cargo de una compañía de teléfonos. Al mismo tiempo, no toda la información de la organización de seguridad es igual de secreta o importante.

[1] Este artículo fue redactado por Fernando Barajas con base en la investigación *Hacia los sistemas de almacenamiento distribuido de información, orientados por la naturaleza de los contenidos*, cuyo responsable es el Dr. Ricardo Marcelín Jiménez con la asistencia de la Dra. Reina Carolina Medina Ramírez y Diego R. Guzmán Santamaría. Los autores colaboran en proyectos de investigación aplicada del Fondo de Información y Documentación para la Industria INFOTEC.

En el estado actual del almacenamiento de la información, existen dos necesidades básicas emergentes que nos empujan a pensar en nuevas estrategias. Por un lado, las organizaciones actuales requieren compartir información más intensa y eficientemente (imaginemos diversos hospitales y consultorios que necesitan compartir historiales clínicos). Por el otro, los equipos de trabajo deben compartir información generada a lo largo del tiempo. En efecto, en un gran proyecto, los distintos participantes aportan conocimiento a la vez que necesitan recuperar los datos generados por sus colaboradores. Un eficiente sistema de almacenamiento solventa esta necesidad al tiempo que echa mano de conceptos comunes que facilitan la recuperación de los documentos de acuerdo con la tarea que cumplen los participantes.

En el primer caso, la simple reestructuración de los procesos no representa ninguna solución, pues las organizaciones han invertido una gran cantidad de recursos para generar sus plataformas de operación; por lo tanto, no se les puede pedir que desechen ese trabajo. La opción, pues, es crear "puentes", traductores informáticos, que hagan compatibles las tareas de las organizaciones actuales que necesiten compartir información. En segundo término, la recuperación de documentos debe mejorar sus capacidades. Así, en un ambiente colaborativo de trabajo, es importante que los miembros puedan acceder a archivos no sólo por el nombre sino por el tema y por palabras clave. De esta forma se garantiza, por ejemplo, que cada actor pueda acceder a archivos que no conoce aún.

Para entenderlo de mejor manera, pensemos en un espacio geográfico como una ciudad. En una urbe es difícil encontrar, digamos, un hospital, puesto

que la ubicación de dicho edificio tiene poca relación con su contenido. Cada construcción, en este caso (casa, comercio, oficina, etc.), corresponde a un tipo de información en un SAD. Si anotamos semánticamente a los edificios y los ordenamos de acuerdo con ello, sin duda sería más fácil ubicarlos en nuestra ciudad. La anotación semántica de un SAD lo define en gran medida, pues con ella es posible dibujar mapas de acceso más eficientes: si yo busco un hospital, el sistema automáticamente me dibujará un mapa de acceso a él, lo mismo si busco un documente en un SAD con determinadas características.

Sin duda, una organización puede obtener grandes beneficios echando mano de los SAD; sin embargo, si además este sistema atiende a la naturaleza de la información y a la manera de distribuirla y acceder a ella, los beneficios se multiplican notablemente. De tal forma, la semantización de sistemas P2P de distribución (*peer to peer*, o entre pares) es un área de especial interés y desarrollo. En pocas palabras, se trata de dibujar mapas para que las terminales de red sepan dónde buscar la información que se les pide.

Existen dos propuestas para este tipo de estructura: por un lado, el uso de índices que estarían acomodados de acuerdo con las anotaciones de los documentos y, por el otro, las búsquedas que toman en cuenta dichos índices, pero también el contenido de los archivos. Así, los resultados incluirían tanto los relacionados por el índice como por el contenido mismo. La búsqueda semántica de documentos puede realizarse por medio de llaves (al modo de un cuadro sinóptico), de manera que el "mapa" semántico se dibuja como un plano cartesiano con varias dimensiones, donde se indica específicamente la ruta a

seguir para encontrar un archivo. Sin embargo, este esquema presenta muchos problemas: se necesitaría un número muy elevado de dimensiones, se requieren muchas copias de un archivo, es difícil expandir mucho las llaves y los resultados no se presentan de manera jerárquica. Para resolver esto se puede pensar no en un "mapa" plano, sino en uno de tres dimensiones, en sustitución de un mapamundi o un globo terráqueo. El espacio se diversifica y puede asociarse con terminales de red (computadoras, servidores), además de que se agrega la noción de distancia.

Un elemento que es importante considerar a la hora de anotar semánticamente un SAD es la heterogeneidad. En efecto, un sistema P2P está compuesto de varios sistemas de usuarios que interactúan a la par y en el mismo nivel. Por ello, son diversos, pues cada sistema guarda su propia lógica interna. Tratar de nivelar los sistemas es una tarea poco menos que imposible, pues, en un SAD de gran volumen, los sistemas tendrían que congelarse y el crecimiento tendría que suspenderse, cosa impensable en un P2P. Por ello, es necesario generar estándares universales que ayuden a los sistemas a interactuar. En ese sentido podemos pensar en "traductores" para los "mapas", es decir, es como si cada participante hablara un idioma diferente y los traductores de mapas ayudaran a cada uno a entender los lugares del mapa en su propia lengua. De esta forma, se respeta la autonomía de cada sistema al tiempo que se posibilita su relación en un SAD.

Otra propuesta interesante es organizar los datos de acuerdo a vecindarios con significado. Imaginemos una red SAD de música que distribuye la información.

En este caso, un usuario comparte determinados archivos musicales y el sistema les asigna un vecindario relativo al género musical. Así, el mapa dibuja vecindarios de música clásica, jazz o rock, al tiempo que genera rutas de acceso de acuerdo a la búsqueda realizada. Los vecindarios pueden ser aún más específicos y contener información jerarquizada, como música clásica de la época barroca. Con ello, si alguien realiza una búsqueda específica, los resultados presentarían una jerarquía similar a la que tienen los vecindarios, y la ruta marcada partirá desde el lugar donde está almacenado lo que se solicita. Este sistema se basa en anotaciones semánticas extendidas sobre redes de almacenamiento, es decir, atiende tanto a la posición del archivo en el sistema como a su contenido.

En resumidas cuentas, un SAD sin anotación semántica está cancelando una gama importante de beneficios. Sin embargo, y a pesar de todo lo anterior, la anotación semántica tiene sus límites, pues no existe hasta ahora una manera de que se hagan inferencias sobre el conocimiento, lo cual quiere decir que una computadora que anota semánticamente y de manera automática un documento pasará por alto muchas implicaciones y sentidos importantes. No será nunca, pues, perfecta. Así que a la hora de que una organización elija un sistema de anotación de SAD, debe considerar el tipo de orden y recuperación que le interesa de acuerdo con la información que maneja: los mapas pueden ser muy exactos, pero no son lo que representan, pues sólo constituyen una imagen. Cada mapa le da importancia a elementos diferentes (calles, ríos, montañas, zonas horarias, etc.) y le resta valor a los demás. Ésta es una razón más para que nuestro almacén se construya y ordene de acuerdo con lo que guardaremos en él.

Si te interesó el artículo, también puedes consultar:

- Investigación *"Hacia los sistemas de almacenamiento distribuido de información, orientados por la naturaleza de los contenidos"*
- Artículos de Divulgación INFOTEC
- Proyectos de Investigación Aplicada en INFOTEC

Ricardo Marcelín Jiménez

r.marcelin.jimenez@gmail.com

Doctor en Ciencias Computacionales por la Universidad Autónoma Metropolitana Unidad Iztapalapa. Miembro del Sistema Nacional de Investigadores Nivel I, otorgado por el CONACYT. Profesor del Área de Redes y Telecomunicaciones, del Departamento de Ingeniería Eléctrica de la Universidad Autónoma Metropolitana, Unidad Iztapalapa. Como investigador, sus intereses son: el almacenamiento distribuido, las redes inalámbricas de sensores y la simulación de eventos discretos. Actualmente, entre otras actividades, colabora en proyectos de investigación en INFOTEC, dirigiendo el proyecto de investigación *"Hacia los sistemas de almacenamiento distribuido de información, orientados por la naturaleza de los contenidos".*