

04 DE OCTUBRE DE 04

Textpreso mezcla de forma más rica los datos científicos

Frustrados porque el volumen y la complejidad de la creciente literatura científica podrían hacer imposible que los investigadores se mantengan al día, investigadores del Instituto Médico Howard Hughes han desarrollado el Textpreso, un nuevo sistema buscador de texto que examina la literatura científica e identifica la información relevante casi tan bien como los curadores de documentos humanos.

El Textpreso fue desarrollado originalmente por los investigadores de HHMI para organizar una base de datos de 7.000 artículos y de 20.000 resúmenes científicos sobre investigación realizada con el gusano redondo *C. elegans*. Los científicos ahora están poniendo el software a disposición de la comunidad científica para organizar las bases de datos que crecen rápidamente con literatura científica en otros organismos y disciplinas.

Los científicos que desarrollaron el Textpreso, Hans-Michael Müller, Eimear Kenny y el investigador del Instituto Médico Howard Hughes, Paul W. Sternberg, todos del Instituto de Tecnología de California, publicaron un artículo sobre el proyecto en número de noviembre 2004 de la revista *Public Library of Science Biology*. El artículo fue publicado en Internet de forma adelantada a la publicación impresa.

"Me encuentro con más tiempo para leer artículos, porque la computadora está examinando los artículos para que yo encuentre los que realmente necesito leer. No esperaba que eso sucediera."

- Paul W. Sternberg

Sternberg dijo que el Textpreso nació a partir de una sensación de frustración que él y sus colegas encontraron al organizar WormBase, una colección de datos científicos sobre el gusano redondo. "Un problema era que teníamos a

biólogos muy altamente entrenados tratando de extraer información de las publicaciones e incorporarlas en un formato uniforme y accesible para la gente”, dijo Sternberg. “Es lento y tedioso, y queríamos acelerarlo y automatizarlo -especialmente para mantener feliz a esta gente talentosa, inteligente y motivada”.

“El otro asunto era que deseamos darles a los biólogos una herramienta de búsqueda de texto más eficaz para permitirles encontrar la información necesaria más eficientemente”, dijo.

Sternberg dijo que el enorme aumento en la complejidad de la literatura científica con el advenimiento de los proyectos genómicos a gran escala hace fundamental la necesidad de sistemas avanzados de recuperación de texto. “Cada año se duplica la cantidad de artículos que necesito leer personalmente”, dijo. “Ahora es cien veces peor que hace apenas algunos años”.

A diferencia de los motores de búsqueda que se basan en palabras claves individuales, el Textpreso utiliza una ontología para organizar la información en una base de datos de texto, dijo Sternberg. La ontología del Textpreso consiste en un catálogo de tipos de objetos y de conceptos que pueden utilizarse para buscar entre las oraciones de los artículos. Estos objetos y conceptos incluyen términos como “gen” o “célula”, pero también a los términos que relacionan a los objetos, tales como “asociación” y “regulación”, o describen a un objeto, tal como “proceso biológico”.

Una vez que se creó la ontología para la literatura del gusano redondo, el programa Textpreso encontró automáticamente esos términos en la base de datos de textos y los marcó utilizando XML, un lenguaje para marcar documentos.

Después de que los documentos se han codificado, los usuarios pueden buscar la base de datos de una manera mucho más sofisticada y útil que se puede realizar utilizando sólo palabras claves, dijo Sternberg. Las búsquedas pueden consistir en preguntas semánticas más significativas que utilicen combinaciones de términos en la ontología para obtener resultados más exactos.

“Cuando se escriben las palabras claves, se encuentra ciertamente mucho material que se desea, pero también mucho que no se desea”, dijo. “Sólo el tipo de categorización de Textpreso permite búsquedas con una mayor precisión, produciendo más información relevante y menos información irrelevante”, dijo.

Para evaluar la exactitud del Textpreso, los investigadores compararon la actuación de Textpreso y de curadores humanos para identificar oraciones en un grupo de artículos de revistas que contenían información sobre la interacción entre dos genes particulares. Textpreso identificó cerca del 62 por

ciento de las oraciones y los curadores humanos encontraron cerca del 71 por ciento.

“En general, encontramos que había cosas que se les escapaban a los seres humanos y cosas que se le escapaban a la máquina”, dijo Sternberg. Sin embargo, dijo, el funcionamiento de ambos era comparable. En pruebas más grandes, se utilizó a Textpreso para buscar información sobre la interacción entre dos genes particulares. Los investigadores encontraron que Textpreso buscaba con mucha más eficiencia.

Sternberg y sus colegas ahora están haciendo que Textpreso esté disponible para organizar y buscar otras bases de datos, incluyendo la literatura científica sobre *Drosophila*, levadura y temas de neurología. La adaptación de Textpreso involucra primero el “desmenuzar” la ontología en categorías que son generales a todas las bases de datos. Entonces, objetos y conceptos particulares de la base de datos se pueden agregar a la ontología, dijo Sternberg. Para permitir el uso interdisciplinario completo del sistema, los investigadores también necesitarán cambiar la estructura subyacente de la base de datos. El equipo cree que el sistema puede extenderse de modo que extraiga automáticamente hechos que, entonces, se pueden colocar en una base de datos tal como WormBase con sólo una revisión mínima de un curador experto.

Sternberg dijo que las revistas comerciales deberían considerar al sistema como ventajoso porque una búsqueda realizada utilizando a Textpreso sólo produciría oraciones individuales relevantes a la búsqueda. Un científico, entonces, obtendría el artículo completo a través de los canales regulares. Inversamente, los científicos podrían utilizar a Textpreso para determinar si los artículos en particular son relevantes antes de acceder a ellos mediante una suscripción.

Personalmente, dijo Sternberg, Textpreso ha hecho que su búsqueda de información científica sea mucho más eficiente. “Me encuentro con más tiempo para leer artículos, porque la computadora está examinando los artículos para que yo encuentre los que realmente necesito leer, y no esperaba que eso sucediera”.