REVISTA BITS DE CIENCIA

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN
Universidad de Chile Nº 1 2008

Dirección

Departamento de Ciencias de la Computación
Av. Blanco Encalada 2120, tercer piso
Santiago, Chile
837-0459 Santiago
www.dcc.uchile.cl
Teléfono: 56-2-978-0652
Fax: 56-2-689-5531
dcc@dcc.uchile.cl

Versión actualizada el 22 de agosto de 2008.

Índice general

Editorial

La Computación es una disciplina que incorpora aspectos tanto científicos como tecnológicos, y eso la hace muy peculiar. Mientras casi cualquier persona tiene acceso a un computador, y hace uso de tecnologías como la Web en forma cotidiana -de hecho la Computación incluso moldea nuestras vidas-, detrás de esa tecnología está la ciencia; aspecto conocido sólo por un grupo muy reducido, normalmente circunscrito a ámbitos académicos y profesionales expertos en la disciplina.

La investigación científica en Computación que sobresale es aquella que tiene una clara conexión con la tecnología, y viceversa. Parafraseando a Donald Knuth, la mejor teoría está inspirada en la práctica, y la mejor práctica es la inspirada en la teoría.

Ambos aspectos, el científico y el tecnológico, se cultivan en nuestro Departamento. Realizamos investigación de primera línea a nivel internacional, en una sorprendente cantidad de áreas. Asesoramos a empresas y al Gobierno de Chile en diversos aspectos tecnológicos, desde certificación de calidad de software hasta cuestiones de estandarización. Formamos profesionales en nuestra Ingeniería y Magíster Profesional; asimismo docentes y científicos mediante nuestros Magíster y Doctorado, impartidos a estudiantes locales y del interior del país, así como latinoamericanos y europeos.

Estamos convencidos que es nuestro rol divulgar los temas científicos y tecnológicos a la comunidad, con el firme propósito de estimular el interés y la discusión sobre los aspectos más contingentes de esta disciplina. Sobre esa certeza concebimos la presente Revista; como un vehículo para llevar artículos, entrevistas y opiniones sobre ciencia y tecnología a los actores sociales más comprometidos con la disciplina: instituciones académicas, gubernamentales, empresariales y, por supuesto, a todos aquellos interesados en la Ciencia de la Computación.

Los invitamos a disfrutar de este primer número, y de todos los que vendrán.

Agosto 2008

Gonzalo Navarro
Director
Departamento de Ciencias de la Computación
Universidad de Chile

Nuestro Departamento

En pocos años y a una velocidad insospechada la computación invadió todos nuestros espacios. Los avances de la informática cambiaron nuestra manera de comunicarnos, de concebir el tiempo y la distancia. Conscientes de que esta revolución tecnológica recién se iniciaba, en 1975 un grupo de académicos visionarios de la Universidad de Chile fundan el Departamento de Ciencias de la Computación (DCC). La misión: ser un centro de excelencia en educación, investigación e innovación en las diversas áreas de la Computación para Chile y el mundo.

Pertenecientes a la Facultad de Ciencias Físicas y Matemáticas, somos responsables de impartir la carrera de Ingeniería Civil en Computación; certificada por la Comisión Nacional de Acreditación (CONAP) por siete años, el máximo período contemplado en nuestro país. Asimismo ofrecemos los programas de posgrado Magíster y Doctorado en Ciencias, ambos con mención Computación; y Magíster en Tecnologías de la Información. Todos acreditados también por la CONAP. Contamos además con el Programa de Educación Continua y Capacitación, destinado a profesionales del área que buscan actualizar sus conocimientos sobre las tecnologías emergentes, y adquirir nuevas destrezas y habilidades.

Congregamos a un prestigioso cuerpo académico de jornada completa, constituido por veinte profesores chilenos y extranjeros del más alto nivel. Todos ellos con magísteres y doctorados obtenidos principalmente en Europa, Norteamérica y Chile. Uno de los imperativos de este selecto grupo es realizar investigación científica en el área, y diseñar sistemas de alta complejidad tecnológica e ingenieril; procesos y producción de conocimiento en los que participan activamente nuestros estudiantes. A nuestros científicos se suman docentes de jornada parcial; profesionales de excelencia técnica y vocación pedagógica que trabajan en la industria de la informática y tecnología. En el Departamento anualmente implementamos exitosas experiencias de investigación y desarrollo de tecnologías lideradas por nuestros investigadores. Como Departamento otorgamos asesorías que contemplan análisis, diseño e implementación de sistemas computacionales; elaboración de bases de licitación, análisis de calidad de procesos de desarrollo de software y plataformas tecnológicas.

En 1987 el DCC se hace cargo del registro de nombres de dominio terminados en ".cl" en acuerdo con IANA, entidad mundial administradora de los nombres de dominio en Internet. Diez años después creamos formalmente NIC Chile que hasta hoy administra los nombres de dominio del país. En 1992 colaboramos en el desarrollo del software de cómputos de sufragios para las elecciones, y en 1993 nos hicimos cargo de él hasta 2000. En 1992 también participamos en la primera conexión a Internet en Chile, y en 1993 instalamos en nuestro edificio el primer servidor Web de Latinoamérica sobre el que construimos el primer sitio Web con información del país. En 1999 creamos el buscador de contenidos en Chile, todoCL. En 2000 implementamos la primera incubadora universitaria de empresas de Chile. Tres años más tarde nuestro programa de Doctorado -creado en 1997- se convierte en el primero de esta área acreditado en el país. Entre 2001 y 2002 diseñamos el sistema de factura electrónica para el Servicio de Impuestos Internos, en uso actualmente, y en 2005 elaboramos la Norma sobre Interoperabilidad de Documentación Electrónica que fija el formato de intercambio de este tipo de documentación perteneciente al gobierno de Chile.

La historia avala nuestra visión científica y tecnológica como Departamento de Ciencias de la Computación, mientras a diario enfrentamos los desafíos de la educación, investigación e innovación. Hoy en el DCC producimos conocimiento y desarrollamos los productos que en el futuro, una vez más, volverán a cambiar nuestra sociedad.

Búsqueda por Similitud en Bases de Datos de Objetos 3D

Benjamin Bustos[**]
Centro de Investigación de la Web,
Departamento de Ciencias de la Computación
Universidad de Chile
bebustos@dcc.uchile.cl

Resumen Los modelos 3D son un tipo importante de dato multimedia, con una amplia gama de aplicaciones prácticas en áreas como la producción industrial, simulación, visualización y entretenimiento. La definición de similitud entre modelos 3D y la implementación de algoritmos de búsqueda por similitud son vitales para la puesta en marcha de bases de datos de objetos 3D. Sin embargo, esto conlleva al mismo tiempo problemas difíciles de resolver. En este artículo se presentan y discuten métodos para implementar la recuperación eficaz de modelos 3D en bases de datos multimedia.

1. Introducción

El desarrollo de métodos eficientes y eficaces de búsqueda para tipos de datos multimedia, como imágenes y video, se ha convertido en un tema de investigación importante debido a la creciente disponibilidad de información audiovisual. Un desarrollo similar se espera para los datos 3D, dado que los modelos 3D son un medio interesante para la difusión y el procesamiento de información en aplicaciones en las áreas de diseño industrial, simulación y entretenimiento, por dar algunos ejemplos. Todas estas aplicaciones tienen en común que las consultas realizadas a la base de datos no son búsquedas exactas (como en las bases de datos tradicionales), sino que son *búsquedas por similitud*, es decir, la consulta es un objeto 3D y uno desea recuperar todos los objetos 3D en la base de datos que sean *geométricamente similares* a la consulta. A éste tipo de búsqueda se le conoce también como *búsqueda por contenido*.

Benjamín Bustos es profesor Asistente del DCC. Doctor en Ciencias Naturales de la Universidad de Konstanz, Alemania (2006), Magíster en Computación (2002) e Ingeniero Civil en Computación de la Universidad de Chile (2001). Su principal línea de investigación es Bases de Datos Multimedia.

[**] Parcialmente financiado por el Nucleo Milenio Centro de Investigación de la Web, P04-067-F, Mideplan, y por el Proyecto FONDECYT 11070037.

Si bien existen muchas formas de definir y diseñar modelos 3D, una de las más comunes es a través de una *malla de triángulos*, aunque también existen representaciones basadas en aproximaciones volumétricas o en nubes de puntos. Más información sobre formas de representar objetos 3D se puede encontrar en Campbell y Flynn [8]. Cualquiera de estas representaciones puede utilizarse para realizar búsquedas por similitud en bases de datos 3D. La Figura 1 muestra un ejemplo de un modelo 3D representado como una malla de triángulos.

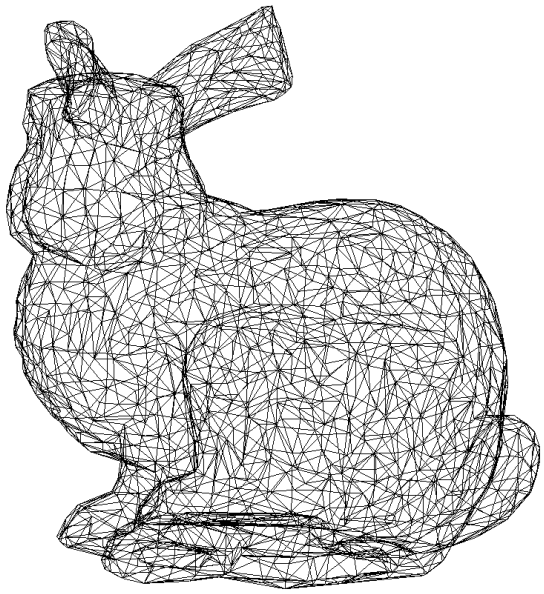

Figura 1. Malla de triángulos de un modelo 3D.

1.1. Aplicaciones

El problema de buscar objetos 3D similares tiene muchas aplicaciones prácticas. Algunos ejemplos son los siguientes:

- En medicina, las tomografías computacionales pueden ser utilizadas para detectar deformaciones de órganos similares a las almacenadas en bases de datos especializadas, lo cual puede ayudar a realizar diagnósticos médicos [9]
- La clasificación estructural es una de las operaciones básicas en biología molecular. Esta clasificación puede realizarse a través de búsquedas por similitud, donde las proteínas y moléculas son modeladas como objetos 3D [1].
- Algunos centros de reportes meteorológicos incluyen "pronósticos de polen" para prevenir y ayudar a las persona alérgicas a los diferentes tipos de polen. Ronneberger et al. [11] desarrollaron un sistema de reconocimiento de patrones que clasifica el polen a partir de información volumétrica (3D).

- Una base de datos de objetos 3D puede servir de apoyo a herramientas de tipo CAD (*Computer Aided Design*). Por ejemplo, partes estándar en una industria manufacturera pueden ser modeladas como objetos 3D. Al diseñar un nuevo producto, compuesto por muchas partes pequeñas, se puede intentar reemplazar algunas de sus partes por piezas estándar si éstas son similares, reduciendo así los costos de producción.

1.2. Búsquedas por similitud

El método estándar para realizar búsquedas por contenido en bases de datos multimedia está basado en el uso de los llamados "vectores característicos". En este método se extraen atributos numéricos desde el objeto multimedia, los cuales describen características importantes del objeto. Estos atributos numéricos se utilizan para construir un vector, el *vector característico* del objeto, el cual es una representación del objeto multimedia original. Estos vectores característicos poseen habitualmente una alta dimensionalidad (del orden de cientos de dimensiones). De esta forma, el problema de comparar objetos multimedia complejos se reduce a buscar puntos cercanos en un espacio vectorial. Como medida de cercanía (similitud) de los vectores característicos se pueden utilizar diversas funciones de distancia, como por ejemplo la distancia Euclidiana o la distancia Manhattan. Es decir, mientras más cercanos estén dos vectores en el espacio vectorial, más similares son.

Hay dos consultas típicas de búsqueda por similitud. Por una parte, la *consulta por rango* retorna todos aquellos objetos de la base de datos que se encuentren a lo más a una cierta distancia ε (definida por el usuario) del objeto de consulta. Por otra parte, la *consulta por k-vecinos más cercanos* retorna los k objetos de la base de datos más cercanos al objeto de consulta.

2. Extracción de características desde objetos 3D

Existen muchos métodos distintos para describir objetos 3D como vectores característicos [5]. En esta sección se revisarán requisitos básicos para todos los descriptores 3D, así como un modelo general de extracción de caraterísticas.

2.1. Requerimientos de invarianza

Considerando el método basado en descriptores, se pueden definir varios requerimientos que estos descriptores deben cumplir. Un buen descriptor 3D debe ser *invariante* a cambios en la orientación (traslación, rotación y reflexión) y en la escala del modelo 3D. Esto significa que el sistema de búsqueda debe ser capaz de recuperar objetos 3D geométricamente similares con orientaciones y tamaños diferentes. Además, un buen descriptor 3D debe ser *robusto* con respecto a pequeños cambios en el nivel de detalle, geometría y topología de los modelos, es decir, el vector característico de un modelo 3D no debiera variar mucho frente a pequeños cambios en el modelo original.

Las propiedades de invarianza y robustez se pueden obtener en forma implícita por descriptores que consideran propiedades relativas de los objetos 3D, por ejemplo, la distribución de la curvatura de la superficie del objeto. Para otros descriptores, estas propiedades se pueden obtener a través de un *proceso de normalización* que transforma los objetos de manera que quden representados en un sistema de referencia canónico, donde las distancias y las direcciones son comparables entre

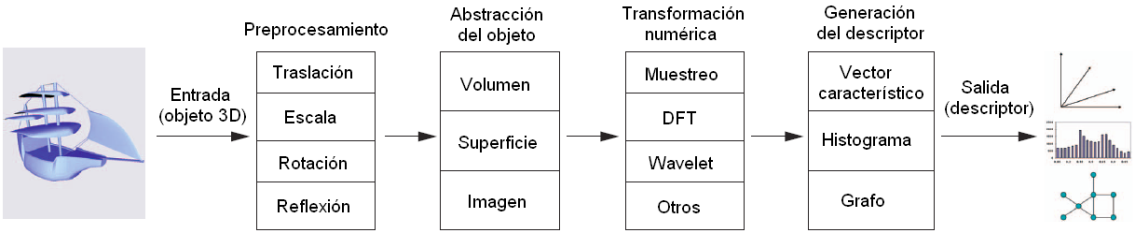

Figura 2. Modelo del proceso de extacción de descriptores 3D.

modelos 3D distintos. El método predominante para encontrar dicho sistema de referencia canónico está basado en el ánalisis de componentes principales (PCA) [10,12]. La idea de este método es alinear el objeto 3D considerando su centro de masa y sus ejes principales. El objeto se traslada en el espacio de forma que su centro de masa coincida con el origen del sistema de coordenadas (invarianza a traslaciones). Luego es rotado alrededor del origen de forma que los ejes x, y y z coincidan con las tres componentes principales del objeto (invarianza a rotaciones). La invarianza a reflexiones se puede obtener mediante un test basado en momentos, y la invarianza a escalamientos se puede obtener mediante el escalamiento del objeto por un factor canónico.

2.2. Modelo del proceso de extracción de características

Los distintos métodos de extracción de características para objetos 3D se pueden representar en un modelo general del proceso [6]. Este proceso se compone de varias etapas (ver Figura 2). Para un objeto 3D dado, usualmente representado como una malla de triángulos, se realiza primero un preprocesamiento (normalización) para obtener las propiedades de invarianza y robustez. Luego se realiza una abstracción del objeto 3D, de forma de caracterizarlo a partir de propiedades en su superficie, propiedades volumétricas o a partir de proyecciones en 2D (imágenes) del objeto. A continuación se puede realizar un análisis numérico de la forma 3D (por ejemplo un muestreo, o aplicar la transformada discreta de Fourier o la transformada Wavelet, etc.), y de este resultado se extrae finalmente el vector característico.

i. *Preprocesamiento.* Si se requiere, se normaliza el objeto 3D para obtener invarianza a rotaciones, traslaciones, escalamientos y reflexión.
ii. *Tipo de abstracción.* Hay tres tipos distintos: *superficie* (se miden características de la superficie del objeto), *volumen* (se miden características del volumen que ocupa el objeto 3D en el espacio), *imagen* (se toman proyecciones 2D del objeto en diferentes direcciones y se analizan posteriormente).
iii. *Transformación numérica.* Las características principales de la malla de triángulos se pueden capturar numéricamente usando distintos métodos. Por ejemplo, grillas de celdas o conjuntos de imágenes se pueden transformar utilizando la transformada de Fourier, o se pueden realizar muestreos sobre la superficie del objeto.
iv. *Generación del descriptor.* Puede ser de tres tipos: un *vector característico* (cuyas coordenadas la conforman los atributos numéricos extraídos del objeto), un *histograma* (que resume alguna característica medida del objeto) o un *grafo* (que puede representar la estructura del objeto 3D).

3. Calidad de la recuperación

En Bustos et al. [6] se describen los 16 descriptores 3D a evaluar en la comparación experimental.

3.1. Evaluación experimental

La base de datos utilizada para realizar los experimentos consiste en 1,838 objetos 3D recopilados de Internet[1]. Una parte de este conjunto (472 objetos) fue clasificada manualmente por similitud geométrica en 55 clases distintas de modelos. El resto de los objetos fue considerado como "no clasificado". Cada objeto perteneciente a una de estas 55 clases se utilizó en la evaluación experimental como objeto de consulta, y los objetos pertenecientes a su misma clase son los considerados como relevantes para la respuesta. Para comparar los distintos descriptores, se utilizaron diagramas de *precisión vs. recuperación* [2]. *Precisión* es la fracción de objetos recuperados que son relevantes a la consulta, y *recuperación* es la fracción del total de objetos relevantes recuperados. Todos los diagramas de precisión vs. recuperación se basan en los once niveles estándar de recuperación (0 %, 10 %, ..., 100 %), y se promediaron los resultados sobre todas las consultas en cada nivel de recuperación. Adicionalmente se utilizó la *R-precisión* (un valor escalar), que mide el valor de la precisión cuando el sistema retorna R objetos, donde R es el número de objetos *relevantes* para la consulta.

3.2. Comparación de la eficacia de distintos métodos

La Figura 3 (superior) muestra los resultados obtenidos en la evaluación experimental. El descriptor más eficaz (en promedio) es el denominado *Depth Buffer* (un descriptor basado en proyecciones 2D del objeto 3D), con una R-precisión promedio de 32 %. La diferencia de eficacia entre el mejor y el peor descriptor (*Depth Buffer* y *Shape Spectrum*, respectivamente) es significativa (factor 3×). Sin embargo, la diferencia entre los mejores descriptores es pequeña. Esto significa que en la práctica los mejores descriptores son, en promedio, similares en su eficacia.

La Figura 3 (inferior) muestra la influencia de la dimensión del vector característico en la calidad de la respuesta obtenida (medida como R-precisión). Se observa que la eficacia crece con la dimensión, pero la tasa de mejora disminuye rápidamente a partir de aproximadamente 64 dimensiones para casi todos los descriptores estudiados, alcanzándose un punto de saturación.

3.3. Análisis de los resultados

A partir de los resultados obtenidos, se puede concluir que los mejores descriptores 3D (en promedio) son aquellos basados en proyecciones del objeto original (e.g., *Depth Buffer*, *Silhoutte*, *Rays-SH*). También obtuvieron buenos resultados algunos descriptores que extraen características volumétricas de los objetos 3D (e.g., *Voxel*, *3DDFT*). Los descriptores basados en características de la superficie del objeto 3D mostraron en general una eficacia baja. Todos los descriptores implementados mostraron ser robustos con respecto al nivel de detalle de los objetos.

Sin embargo, también se observó una varianza alta con respecto a la eficacia de los descriptores cuando se compararon los resultados entre distintas clases de objetos. Por ejemplo, para la clase "autos Fómula 1" el mejor descriptor fue *Depth Buffer* (el mejor descriptor en promedio), mientras

[1] La base de datos está disponible en `http://merkur01.inf.uni-konstanz.de/CCCC/`.

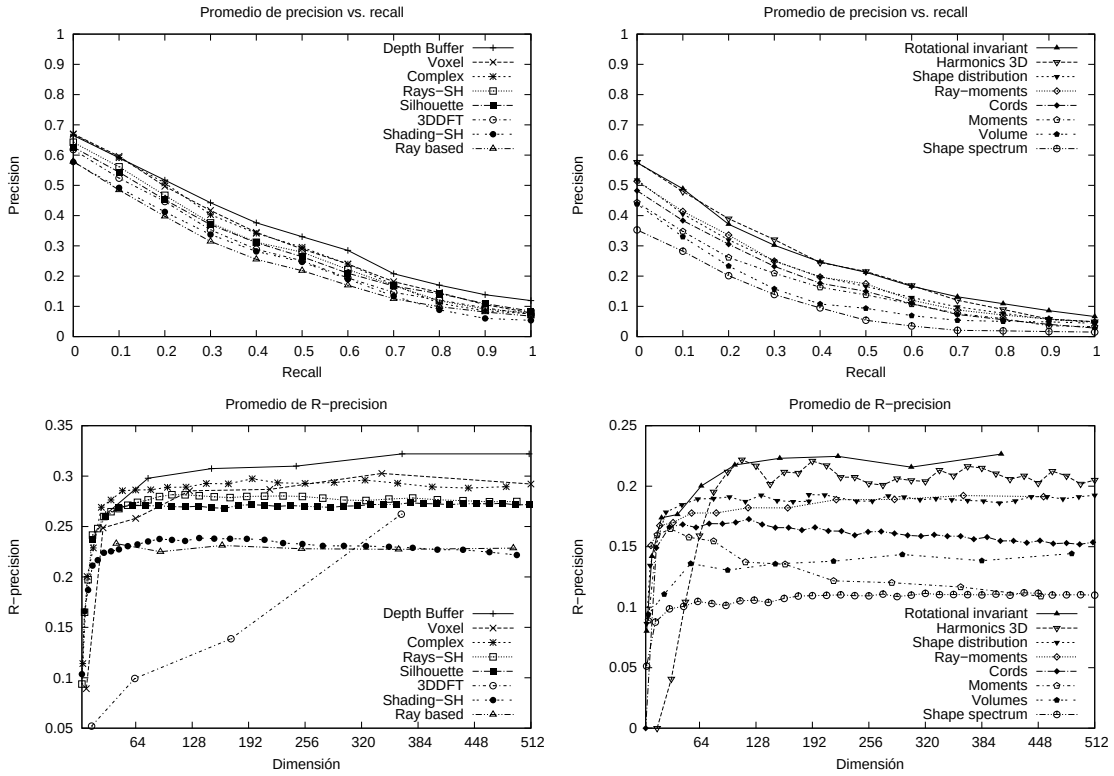

Figura 3. (Figuras superiores) *Precision vs. recall* para 16 descriptores 3D. (Figuras inferiores) *R-precision* en función de la dimensión para los mismos descriptores 3D.

que para la clase "animales marinos" el mejor descriptor fue Silhouette, y su eficacia fue casi el doble mejor que con *Depth Buffer*. Inclusive, para una clase muy particular, "figuras humanas", el mejor descriptor fue *Shape Spectrum*, que resultó ser el peor en promedio.

Aparte de esta última notable excepción, no fue posible encontrar una correlación fuerte entre la geometría del objeto 3D y el mejor descriptor para dicha geometría. Por último, también se observó que para todos los descriptores estudiados se alcanza un punto de saturación a partir de una cierta dimensión, por lo que no es posible mejorar la eficacia de la búsqueda añadiendo más coordenadas a los vectores característicos una vez alcanzado este punto de saturación.

4. Combinando descriptores

Una forma de superar las limitaciones presentadas en la Sección 3.3 es utilizando combinaciones estáticas o dinámicas de descriptores 3D. La idea es combinar distintos descriptores para describir un objeto 3D, dado que los distintos descriptores capturan características distintas del objeto y por lo tanto la calidad de la respuesta podría mejorar al considerar toda esta información en conjunto al realizar una búsqueda. La Figura 4 muestra un ejemplo de tres búsquedas con el mismo objeto de

consulta (un modelo 3D de un auto Fórmula 1), utilizando primero dos descriptores por separado y luego utilizando una combinación simple de ellos. Las consultas con los descriptores utilizados independientemente retornan algunos objetos no relevantes (marcados en la figura), en cambio al combinar ambos descriptores sólo se obtienen objetos relevantes en la respuesta.

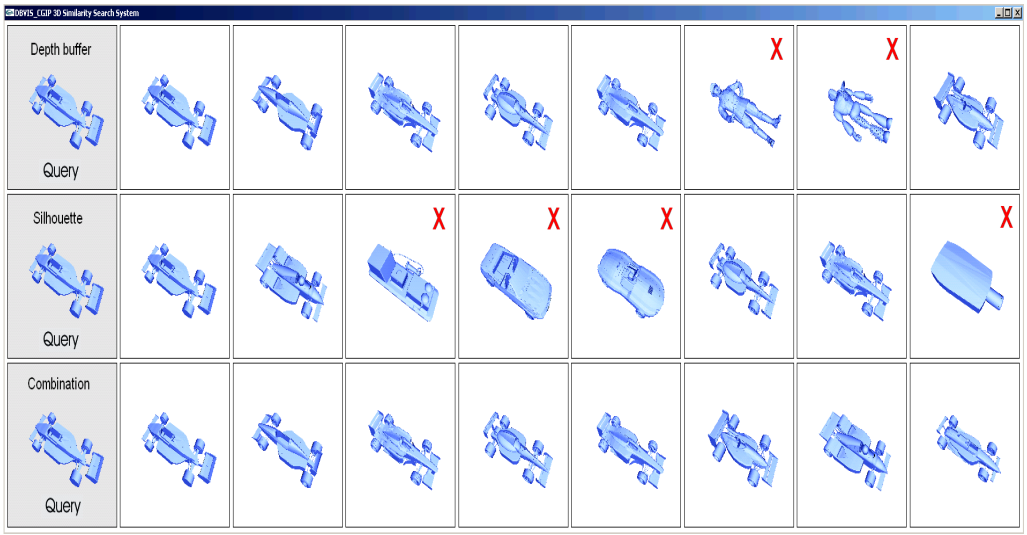

Figura 4. Comparación de dos búsquedas con descriptores independientes y su combinación estática.

4.1. Combinaciones estáticas

Una forma simple de implementar las combinaciones de descriptores 3D es simplemente sumando las distancias obtenidas con cada una de ellas, para luego utilizar este valor para generar el ranking final de la respuesta (éste fue el método utilizado para producir la Figura 4). Este método se denomina *combinación estática*, ya que se utiliza la misma combinación de descriptores para realizar todas las consultas.

Se probaron experimentalmente todas las posibles combinaciones estáticas de dos o más descriptores (hasta combinaciones de 10 descriptores), y el mejor resultado se obtuvo combinando los 5 mejores descriptores según los resultados presentados en la Figura 3. Con esta combinación estática se obtiene un aumento significativo en la calidad de las búsquedas por similitud. La Figura 5 muestra el diagrama de precisión vs. recuperación comparando el mejor descriptor en promedio con la mejor combinación de descriptores. La figura muestra que la combinación no solo aumenta el valor de la precisión para cada punto de recuperación, sino que también el valor de la R-precisión aumenta de 0,32 a 0,42, es decir, un aumento relativo de un 30 % (para dar una idea, aumentos relativos de la R-precisión de un 10 % son significativos).

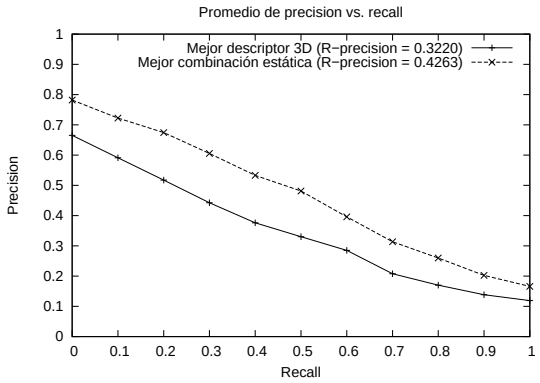

Figura 5. Comparación entre el mejor descriptor 3D (*Depth Buffer*) y la mejor combinación estática de descriptores.

4.2. Combinaciones dinámicas

Si bien se pueden obtener buenos resultados utilizando combinaciones estáticas de descriptores 3D, al analizar los resultados se llega a una conclusión similar que la obtenida con descriptores independientes: la mejor combinación a utilizar no siempre es la que funciona mejor en promedio, sino que nuevamente depende del objeto de consulta. El problema es que para algunos objetos de consulta un descriptor en particular puede ser muy útil, pero para otras consultas puede ser irrelevante.

Para resolver este problema, se propone el uso de *combinaciones dinámicas* de descriptores. En vez de fijar *a priori* la combinación a utilizar, ésta se elige dependiendo del objeto de consulta (que resulta de hacer una selección de descriptores a combinar). Alternativamente, se pueden utilizar todos los descriptores disponibles y lo que se asigna dinámicamente es el *peso* que tendrá cada descriptor en la suma de distancias. Si se decide que un descriptor es "malo" para una consulta dada, su peso debiera ser cercano a 0, en caso contrario, si es un "buen" descriptor para el objeto de consulta dado, su peso debiera ser cercano a 1.

El cálculo de los pesos se puede realizar utilizando una *base de datos de entrenamiento*. La hipótesis es que si la respuesta que retorna un descriptor en esta base de entrenamiento es *coherente* (i.e., los objetos retornados se parecen entre sí), entonces el descriptor se considera "bueno" para la consulta, y se considera "malo" en caso contrario. El método conocido como *entropy impurity* [4] está basado en esta hipótesis.

La figura 6 muestra los resultados obtenidos utilizando combinaciones dinámicas de descriptores, tanto con el método de selección como con el método *entropy impurity* (k es un parámetro del método). Los resultados muestran que el método más eficaz es el método *entropy impurity*, es decir, lo mejor es considerar todos los descriptores disponibles en el sistema de búsqueda, asignándoles su peso correspondiente al momento de realizar una consulta. Con este método, el valor de la R-precisión aumenta de 0,32 (mejor descriptor) a 0,46, es decir, un aumento relativo de un 43 %.

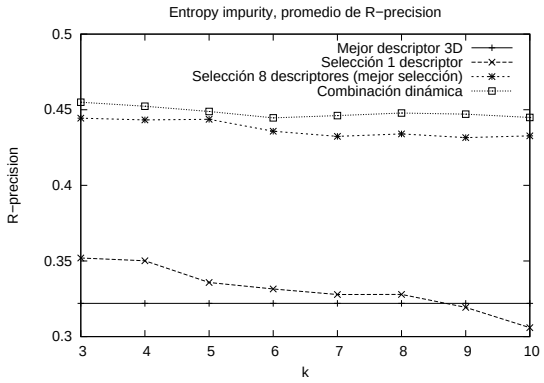

Figura 6. Comparación entre el mejor descriptor 3D (*Depth Buffer*), la mejor selección dinámica de descriptores y la mejor combinación dinámica de descriptores.

5. Problemas abiertos

Existen aún una serie de problemas abiertos relacionados con la recuperación de objetos 3D [3]. Los métodos actuales de búsqueda por similitud se enfocan principalmente en los aspectos geométricos de los modelos 3D, siendo ignorados otros atributos presentes en muchas bases de datos de objetos 3D, como por ejemplo el color, material y textura del objeto 3D. Otro problema abierto importante es el desarrollo de modelos para realizar búsquedas por similitud parcial en bases de datos de objetos 3D. En este tipo de búsqueda se desean recuperar los objetos de la base de datos que tengan alguna sección similar (y no necesariamente el modelo completo) al objeto de consulta, lo cual lo hace un problema mucho más complicado que el problema de búsqueda por similitud global descrito en este artículo. Por último, es tema de investigación actual desarrollar métodos eficientes de búsqueda por similitud, especialmente en el caso cuando se utilizan combinaciones dinámicas de descriptores [7].

Referencias

1. M. Ankerst, G. Kastenmüller, H.-P. Kriegel, and T. Seidl. Nearest neighbor classification in 3D protein databases. In *Proc. 7th International Conference on Intelligent Systems for Molecular Biology*, pages 34–43. AAAI Press, 1999.
2. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.
3. B. Bustos, D. Fellner, S. Havemann, D. Keim, D. Saupe, and T. Schreck. Foundations of 3D digital libraries: Current approaches and urgent research challenges. In *Proc. 1st International Workshop on Digital Libraries Foundations (DLF'07)*, pages 7–12. DELOS Network of Excellence on Digital Libraries (www.delos.info), 2007.
4. B. Bustos, D. Keim, D. Saupe, T. Schreck, and D. Vranić. Using entropy impurity for improved 3D object similarity search. In *Proc. IEEE International Conference on Multimedia and Expo (ICME'04)*, pages 1303–1306. IEEE, 2004.
5. B. Bustos, D. Keim, D. Saupe, T. Schreck, and D. Vranić. Feature-based similarity search in 3D object databases. *ACM Computing Surveys*, 37(4):345–387, 2005.

6. B. Bustos, D. Keim, D. Saupe, T. Schreck, and D. Vranić. An experimental effectiveness comparison of methods for 3D similarity search. *International Journal on Digital Libraries, Special issue on Multimedia Contents and Management in Digital Libraries*, 6(1):39–54, 2006.
7. B. Bustos and T. Skopal. Dynamic similarity search in multi-metric spaces. In *Proc. 8th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR'06)*, pages 137–146. ACM Press, 2006.
8. R. Campbell and P. Flynn. A survey of free-form object representation and recognition techniques. *Computer Vision and Image Understanding*, 81(2):166–210, 2001.
9. D. Keim. Efficient geometry-based similarity search of 3D spatial databases. In *Proc. ACM International Conference on Management of Data (SIGMOD'99)*, pages 419–430. ACM Press, 1999.
10. E. Paquet, A. Murching, T. Naveen, A. Tabatabai, and M. Rioux. Description of shape information for 2-D and 3-D objects. *Signal Processing: Image Communication*, 16:103–122, 2000.
11. O. Ronneberger, H. Burkhardt, and E. Schultz. General-purpose object recognition in 3D volume data sets using gray-scale invariants - classification of airborne pollen-grains recorded with a confocal laser scanning microscope. In *Proc. 16th International Conference on Pattern Recognition*, volume 2, pages 290–295. IEEE Computer Society, 2002.
12. D. Vranić, D. Saupe, and J. Richter. Tools for 3D-object retrieval: Karhunen-Loeve transform and spherical harmonics. In *Proc. IEEE 4th Workshop on Multimedia Signal Processing*, pages 293–298, 2001.

Buscando en la Web

Gonzalo Navarro
Centro de Investigación de la Web
Departamento de Ciencias de la Computación
Universidad de Chile
gnavarro@dcc.uchile.cl

Se dice que los más jóvenes no tienen idea de cómo era buscar información antes que que existiera la Web. Eso es sólo parcialmente cierto. Los menos jóvenes tampoco recordamos gran cosa. Nos resulta un ejercicio de imaginación muy difícil recordar cómo vivíamos cuando, ante cualquier consulta, desde cultural hasta de entretenimiento, no podíamos escribir un par de palabras en nuestro buscador favorito y encontrar inmediatamente montañas de información, usualmente muy relevante.

Gonzalo Navarro es Profesor Titular del DCC, donde obtuvo su doctorado en 1998. Actualmente dirige el Centro de Investigación de la Web (CIW). Sus principales áreas de interés son: algoritmos y estructuras de datos, búsqueda en texto, búsqueda por similitud, y compresión. Ha escrito un libro de búsqueda en texto, y unos 200 artículos en libros, revistas, y congresos internacionales. Ha presidido el Comité de Programa de 6 congresos internacionales y creado el primer congreso en búsqueda por similitud (SISAP).

Para operar este milagro no basta con Internet. Ni siquiera basta con la Web. El ingrediente imprescindible que se necesita son los *buscadores* o *máquinas de búsqueda*. Estos buscadores, cuyos representantes más conocidos hoy en día son probablemente *Google*, *Yahoo!* y *Microsoft MSN*, son los que conocen en qué páginas de la Web aparecen qué palabras (y saben bastante más). Sin un buscador, deberíamos conocer las direcciones Web de todos los sitios de bibliotecas, o de turismo, o de cualquier tema que nos pudiera interesar, y los que no conociéramos sería como si no existieran. En un sentido muy real, los buscadores *conectan* la Web, pues existen grandes porciones de la Web a las que no se puede llegar navegando desde otra parte, a menos que se use un buscador. No es entonces sorprendente que casi un tercio del tiempo que los usuarios pasan en Internet lo dediquen a hacer búsquedas.

Esto nos da una primera idea del gigantesco desafío tecnológico y científico que supone desarrollar un buscador. Debemos resolver cuestiones básicas como ¿qué páginas debería conocer un buscador? ¿qué debería almacenar de esas páginas? ¿qué tipo de preguntas debería aceptar? ¿qué debería responder a esas preguntas? ¿cómo debería mostrar la información? Y ésas son sólo las preguntas más elementales.

Para ordenar la discusión comencemos mostrando la arquitectura típica de una máquina de búsqueda, en la figura 1. Los cuadrados de bordes duros indican procesos, y los de bordes suaves información almacenada. Las flechas representan flujo de información.

En el *crawling* se recolectan páginas de la Web, ya sea nuevas o actualizadas. El proceso de *parsing* es el que extrae los enlaces que parten de las páginas leídas y realimenta el crawling con

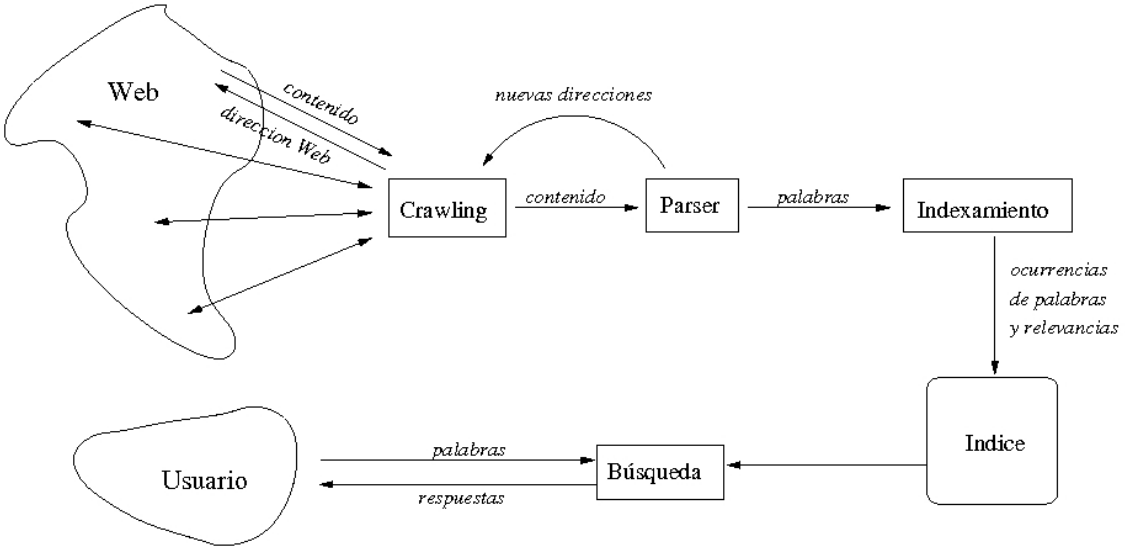

Figura 1. Arquitectura típica de una máquina de búsqueda Web.

nuevas direcciones para visitar, mientras que alimenta al indexador con las páginas depuradas (es decir, sin información irrelevante para el indexamiento). El *indexamiento* almacena en el *índice* la información sobre qué palabras aparecen en qué páginas, junto con una estimación de la importancia de tales ocurrencias. La *búsqueda* usa el índice para responder una consulta, y luego presenta la información al usuario para que éste navegue por ella.

1. Crawling: ¿qué páginas debería conocer un buscador?

Se llama *crawling* al procedimiento de visitar páginas para ir actualizando lo que el buscador sabe de ellas. Un *crawler* es un programa que corre en la máquina del buscador y que va solicitando a distintos computadores de Internet que le transfieran el contenido de las páginas Web que él les indica. Para estos computadores un crawler es prácticamente lo mismo que un humano que visitara sus páginas: debe enviarle el contenido de la página solicitada.

¿Qué páginas debería conocer un buscador? ¡Es tendador responder que todas! Pero lamentablemente esto no es posible. La Web cambia demasiado seguido: un porcentaje alto de las páginas cambia de un mes a otro, y aparece un porcentaje importante de páginas nuevas. Internet no es lo suficientemente rápida: se necesitan meses para transmitir todas las páginas de la Web al buscador. Es simplemente imposible mantener una foto actualizada de la Web ¡Ni siquiera se puede explorarla al ritmo al que va creciendo! La foto que almacena un buscador es siempre incompleta y sólo parcialmente actualizada. No importa cuántas máquinas usemos para el buscador. Los mayores buscadores hoy en día ni se acercan a cubrir la mitad de la Web.

Querer mantener una foto de la Web al día puede compararse con querer estar al tanto de todo lo que ocurre en todas partes del mundo, hasta los menores detalles locales, mediante la continua lectura del diario. Van ocurriendo más novedades de las que es posible ir leyendo. Podemos

pasarnos todo el tiempo leyendo detalles insignificantes y perdiéndonos los hechos más importantes, o podemos tener una política más inteligente de seleccionar las noticias más relevantes, y postergar (tal vez para siempre) la lectura de las menos relevantes. Esto es aún peor si consideramos la llamada *Web dinámica*, formada por páginas que se generan automáticamente, a pedido (por ejemplo al hacer una consulta al sitio de una línea aérea), y que son potencialmente infinitas.

Un tema fundamental en un buscador es justamente el de decidir qué páginas debe conocer, y con cuánta frecuencia actualizar el conocimiento que tiene sobre cada página. Un crawler comienza con un conjunto pequeño de páginas conocidas, dentro de las cuales encuentra enlaces a otras páginas, que agrega a la lista de las que debe visitar. Rápidamente esta lista crece y es necesario determinar en qué orden visitarlas. Este orden se llama "política de crawling". Algunas variables relevantes para determinar esta política son la importancia de las páginas (debería actualizar más frecuentemente una página que es más importante, lo que puede medirse como cantidad de veces que la página se visita, o cantidad de páginas que la apuntan, o frecuencia con que se buscan las palabras que contiene, etc.), y la frecuencia de cambio de las páginas (debería revisitarse más frecuentemente una página que cambia más seguido), entre otras.

2. Indexamiento: ¿qué debería almacenarse de las páginas?

El *indexamiento* es el proceso de construir un *índice* de las páginas visitadas por el crawler. Este índice almacena la información de manera que sea rápido determinar qué páginas son relevantes a una consulta.

¿No basta con almacenar las páginas tal cual, para poder buscar en ellas después? No. Dados los volúmenes de datos involucrados (los mayores buscadores indexan hoy en día miles de millones de páginas, que ocupan varios terabytes), es imposible recorrer una a una todas las páginas almacenadas en un buscador para encontrar cuáles contienen las palabras que le interesan al usuario. ¡Esto demoraría horas o días para una sóla consulta!

El buscador construye lo que se llama un *índice invertido*, que tiene una lista de todas las palabras distintas que ha visto, y para cada palabra almacena la lista de las páginas donde ésta aparece mencionada. Con un índice invertido, las consultas se pueden resolver mediante la búsqueda de las palabras en el índice y el procesamiento de sus listas de páginas correspondientes (intersectándolas, por ejemplo). La figura 2 ilustra un índice invertido.

Los buscadores grandes deben procesar hasta mil consultas por segundo. Si bien este trabajo puede repartirse entre varios computadores, la exigencia sigue siendo alta. El mayor costo para responder una consulta es el de leer de disco las listas de páginas apuntadas por el índice invertido. Es posible usar técnicas de compresión de datos para reducir el espacio en que se representan estas listas. Con esto se logra ganar espacio y velocidad simultáneamente. Pueden hacerse también otras cosas, como precalcular las respuestas a las consultas más populares.

3. Búsqueda: ¿qué preguntas debería responder, y cómo?

Hemos estado considerando que el usuario escribe algunas palabras de interés y el buscador le da la lista de las páginas donde aparecen estas palabras. La realidad es bastante más complicada. Tomemos el caso más elemental, de una consulta por una única palabra. Normalmente hay millones de páginas que contienen esa palabra, y está claro que el usuario no tiene la menor posibilidad

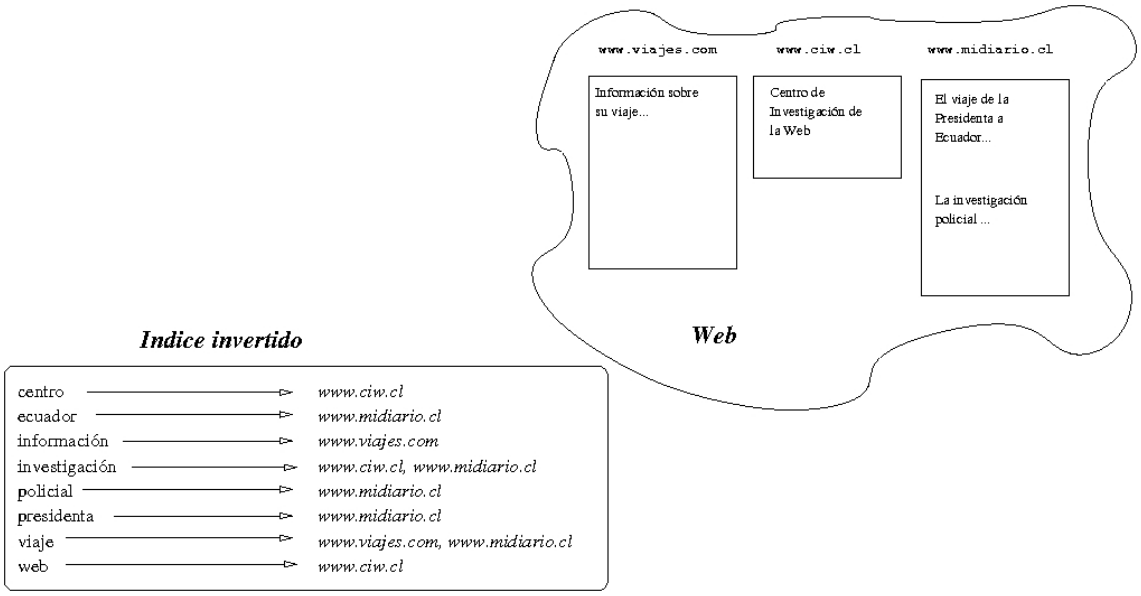

Figura 2. Ejemplo de un índice invertido para tres páginas Web.

de examinarlas todas para ver cuáles satisfacen su necesidad de información. De alguna manera el buscador debe *ordenar* las respuestas por su supuesta *relevancia* a la consulta.

Existen muchas formas de calcular esta relevancia, que dan lugar a mejores o peores heurísticas. Por ejemplo, uno puede considerar que una página donde la palabra aparece varias veces es más relevante que otra donde aparece una vez. Pero si la palabra aparece más veces en una página que es mucho más larga que otra, entonces tal vez la palabra no sea tan importante en esa página. También uno puede considerar cuán importante es la página en sí (por ejemplo si es muy visitada, o muy apuntada por otras). Los buscadores utilizan fórmulas matemáticas para calcular relevancia que tienen en cuenta estos aspectos.

Existen técnicas más sofisticadas, por ejemplo llevar información de cómo se comportaron otros usuarios cuando hicieron esta misma consulta (por ejemplo, el buscador puede saber que la gran mayoría de los usuarios que buscaron "mp3" terminaron yendo a ciertos sitios específicos). Esto se llama *minería de consultas* y es extremadamente útil para dar buenas respuestas a consultas que no dicen mucho. También puede usarse información posicional, por ejemplo si la palabra aparece en el título de la página o de los enlaces que la apuntan, puede ser más relevante que si aparece cerca del final.

La situación se complica cuando la consulta tiene varias palabras, donde algunas pueden ser más importantes que otras. Normalmente las ocurrencias de palabras que aparecen en muchos documentos, como los artículos y preposiciones, son poco importantes porque no sirven para discriminar. Para peor, sus listas de ocurrencias en los índices invertidos son muy largas, ocupando espacio inútil. Por ello muchos buscadores las omiten de sus índices (intente buscar "and" en su buscador favorito). La forma de combinar el peso de las distintas palabras da lugar también a mejores o peores heurísticas. Por ejemplo los buscadores en la Web normalmente muestran sólo

páginas donde aparecen todos los términos, como una forma de eliminar respuestas irrelevantes. Asimismo, los mejores dan preferencia páginas donde las palabras aparecen cercanas entre sí.

La verdad es que en la Web hay mucha, mucha más información de la que se puede obtener mediante la búsqueda de documentos que contengan ciertas palabras. Esta limitación se debe a que no es fácil implementar búsquedas más sofisticadas a gran escala. Conseguir responder consultas más complejas a escala de la Web es un tema actual de investigación. Algunos ejemplos son:

1. Buscar por contenido en fotos, audio o video. Imagínese mostrar una foto de su promoción y poder encontrar otras fotos de las mismas personas en la Web, incluso sin recordar sus nombres. O tararear una parte de una melodía (incluso con errores) y encontrar el mp3 para poder bajarlo. Existen técnicas para hacer esto, pero no a gran escala. Los buscadores ofrecen búsqueda de fotos, pero basada en palabras que se encuentran asociadas a las fotos durante el crawling.
2. Hacer preguntas complejas que se pueden inferir de la Web. Por ejemplo preguntas como ¿cuál es la farmacia más cercana que venda un antigripal a un precio inferior a $ 3.000? y ¿qué universidades dictan una carrera de Diseño Gráfico de 5 años en la Región Metropolitana? Responder este tipo de preguntas requiere normalmente de cierta cooperación de quien escribe las páginas.
3. Hacer consultas con componente temporal, como ¿qué ocurrió con el seguimiento en los medios de comunicación a las consecuencias de la guerra del Golfo en los meses siguientes a su finalización? Esto requiere llevar una cuenta histórica de los contenidos de la Web a lo largo del tiempo.

4. Interacción con el Usuario: ¿cómo presentar la información?

Ya vimos que las respuestas que se muestran al usuario son sólo una mínima parte de las que califican. Los buscadores normalmente presentan una lista de las primeras páginas según el orden que han hecho en base a la consulta. En esta lista se indica la dirección de la página (para que el usuario pueda visitarla con un click) y usualmente el *contexto* del texto donde las palabras aparecen. Esto ayuda al usuario a saber rápidamente si las palabras aparecen en la forma que las esperaba.

Poder mostrar un contexto requiere que el buscador no almacene sólo el índice invertido, sino también el contenido completo de las páginas que indexa, de modo de poder mostrar un pasaje donde aparecen las palabras de la consulta. Si bien el espacio es barato, esto es un requerimiento bastante exigente, pues el buscador debería tener suficiente almacenamiento para duplicar toda la Web en sus discos. Para reducir el espacio, el buscador puede evitar almacenar las imágenes, por ejemplo. La compresión de datos es también útil para aliviar este problema.

Los buscadores suelen ser lo suficientemente buenos como para que, un gran porcentaje de las veces, lo que busca el usuario esté entre las primeras respuestas que ofrece. De todos modos, es posible pedirles que entreguen el siguiente conjunto de respuestas, y el siguiente, hasta hallar lo que uno busca. La experiencia normal es que, si la respuesta no está en las primeras páginas, es raro que esté más adelante. Es mejor en esos casos reformular la consulta, por ejemplo haciéndola más específica (si se hallaron demasiadas páginas irrelevantes) o más general (si se hallaron demasiadas pocas respuestas). Por ejemplo, en la figura 2, si buscáramos "viaje" encontraríamos tanto la página de la agencia de viajes como la noticia sobre el viaje presidencial. Refinando la consulta

a "viaje Presidenta" tendríamos mejor precisión. Esta iteración es frecuente en las sesiones con los buscadores, y con el tiempo el usuario aprende a formular consultas más exitosas.

Existen formas mucho más sofisticadas de presentar la información, pero nuevamente es difícil aplicarlas a sistemas masivos como la Web. Asimismo suele ocurrir que las interfaces demasiado "inteligentes" resultan ser demasiado complejas para la mayoría de la gente. Incluso los lenguajes de consulta más complejos, donde se puede indicar que las palabras A y B deben aparecer, pero no C, normalmente están disponibles en los buscadores Web, pero se usan muy raramente. La regla en este caso es que la simplicidad es lo mejor.

Referencias

- `www.searchenginewatch.com` es un sitio dedicado a las estadísticas sobre las principales máquinas de búsqueda en la Web.
- `http://www.press.umich.edu/jep/07-01/bergman.html` y `http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/` son dos sitios dedicados a estudiar el crecimiento de la Web, y en general de la cantidad de información disponible en el mundo.
- `www.ciw.cl`
- `www.todocl.cl`

Entrevista a Jens Harding sobre Software Libre

por Claudio Gutierrez, profesor del DCC.

Aclarando Conceptos

Software "libre", "open source", propiedad intelectual, patentes, licenciamiento, etc ¿Puedes hacernos un mapa de estos conceptos para el lector que recién comienza a pensarlos?

Todo el tema parte con la idea de que existe gente que trabaja en temas diferentes a la producción de bienes tangibles. Incluso puede ser gente que trabaja dentro del contexto de producción de bienes, como una fábrica, donde se preocupa de temas como la eficiencia, nuevos productos y otros. Si esa persona logra aumentar la eficiencia en, digamos, un 10 %, es relativamente claro cuál es su aporte, y cuánto es lo que la fábrica estaría dispuesta a pagarle como máximo por su servicio. Por otro lado, cuando alguien diseña un nuevo producto que luego se puede fabricar, ya cuesta más determinar el valor de esa actividad, sobre todo si no hay un mercado que vaya equilibrando el precio entre la oferta y la demanda.

Jens Harding es ex-alumno de pregrado y del programa de doctorado del DCC. En la actualidad se desempeña como profesor del Departamento de Ciencias de la Computación de la Universidad Católica de Chile. Su interés como investigador se centra en el impacto social de las tecnologías de la información. Jens es director del CSol, una iniciativa que busca la masificación del uso de tecnologías de Software Libre en Chile.

Pero el problema no se soluciona mediante la incorporación del mercado, porque existe otro tema: una persona es la que dedica esfuerzo para generar un producto intangible: una idea, una receta, un diseño o un invento. Pero una vez creada, cualquiera puede usarla independiente que se le haya ocurrido a él, o simplemente porque supo de la idea. Entonces, el creador de la idea no es capaz de vender su creación, porque en muchos casos basta con ver los efectos o conocer la lógica para poder copiarla, sin necesidad de que el creador original se lo entregue o enseñe. Una consecuencia de esto es que, si no se toma alguna medida, no van a haber muchos interesados en generar bienes intangibles si además tienen que tener un "day job" para subsistir, y no tienen ninguna garantía de que se les vaya a reconocer su aporte.

Aquí surge el concepto de "propiedad intelectual" o, como lo llaman muchos, de "monopolios artificiales sobre bienes intangibles" (MABI). El objetivo es intentar que la humanidad genere la mayor cantidad de conocimiento y arte para maximizar su calidad de vida. Esto lleva al supuesto que como humanidad tendremos mayor calidad de vida mientras mayor sea nuestro progreso en tecnologías y artes.

Pero resulta que maximizar la cantidad de conocimiento y arte disponible para la humanidad no es fácil, porque implica lograr un sutil equilibrio entre diversos actores en muchas situaciones

diferentes. Es justamente por eso que no existe una “propiedad intelectual” global, sino un cúmulo de derechos específicos que poco tienen que ver uno con otro más que aportar a alcanzar ese máximo en algún contexto específico.

Y dentro de cada uno de esos ámbitos específicos, se intenta equilibrar la necesidad de cada uno de los actores involucrados, que al menos involucra a la humanidad en general, a quienes generan conocimiento o arte, y a quienes pretenden utilizar ese conocimiento o arte. Para dificultar aún más la ecuación, los que utilizan el conocimiento o arte incluyen a los que generan el conocimiento o arte posterior, ya que no es posible crear conocimiento o arte de la nada, sino que de alguna forma nos basamos en todo el conocimiento anterior y aportamos un grano de arena. Thomas Edison hizo famosa una frase en ese contexto: "If I have seen a little further it is by standing on the shoulders of giants".

Las patentes, derecho de autor, secreto industrial, marcas y diseños industriales, así como las denominaciones de origen o indicaciones geográficas, son diversas herramientas que comúnmente se agrupan bajo el paraguas de “propiedad intelectual” o “monopolios artificiales sobre bienes intangibles”, y apuntan a regular un contexto específico. En cada caso, se busca encontrar el equilibrio que permita llegar lo más cerca posible del máximo beneficio para la humanidad, tarea que no es nada de fácil y para lo cual aún, como sociedad, estamos en pleno aprendizaje.

En general, a quien genera un beneficio (conocimiento o arte), llamémosle contribuyente, se le da una especie de exclusividad (monopolio), y la sociedad apoya a ese contribuyente a hacer efectiva esa exclusividad mediante el aparato legal y policial. En algunos casos, se le pide cierto nivel de calidad a la contribución, en otros casos solamente se pide originalidad. Y el ingrediente final para intentar establecer el equilibrio son las limitaciones (por ejemplo, que el beneficio tenga un límite de tiempo) y las excepciones (que en algunos casos específicos, la exclusividad no aplica). Y como los contribuyentes en general son mejores aportando en lo que saben que echando a andar fábricas y líneas de producción o imprimiendo y encuadernando los libros con sus creaciones, pueden cederle todos o algunos de los derechos exclusivos que tienen a otros (a través de un contrato legal llamado licencia), generalmente a cambio de un pago. Con eso, si todo funciona bien, tendremos un equilibrio en el cual todos ganan y tenemos que los buenos contribuyentes pueden vivir muy bien de lo que hacen mejor, los contribuyentes promedio tienen incentivos para mejorar, y a los malos contribuyentes probablemente les vaya mejor dedicándose a otra cosa.

Entiendo que el tema del software libre es un tema muy moderno. Esta discusión no se daba (ni tenía sentido), digamos, hace un siglo ¿Puedes explicarnos la esencia de las nociones involucradas?

Todo el tema de los monopolios artificiales sobre bienes intangibles, si bien no se le puede calificar de nuevo, sigue estando en evolución. El caso del software es particularmente nuevo, y hubo que decidir si inventar una herramienta legal nueva para este tema específico, o utilizar una ya existente. De las herramientas ya existentes se tiene por un lado el derecho de autor, y por otro el de las patentes de invención.

Las patentes de invención se le entregan a alguien que ha hecho un aporte importante, no trivial y que tiene aplicación industrial. Lo que se protege en este caso es la idea, y no la forma de expresarla. Si pensamos en una receta para hacer galletas, en este caso lo que se protegería sería el proceso que se usa para hacer las galletas, pero no la receta. De hecho, la descripción tiene que ser pública y suficiente para poder reproducir el proceso por cualquiera con conocimiento del estado del arte, así que yo podría fotocopiar la descripción y regarlarla o venderla a quien quiera. A cambio, cualquiera tiene prohibido utilizar ese proceso, incluso si no lo ha visto ni sabe

de su existencia, salvo que tenga un permiso (licencia) por parte del poseedor de la patente. Para obtener una patente es necesario pasar por un largo y caro proceso que valida que se cumpla con los requisitos, y después de una cierta duración se acaba el beneficio y no es necesario pedir permiso.

El derecho de autor funciona al revés: se protege una expresión, pero no la idea. En el mismo caso de las recetas de galletas, ahora protegidas por derecho de autor y no por patentes: yo podría perfectamente aprenderme la receta y escribirla en mis propias palabras, o seguir la receta al pie de la letra y hacer galletas, y ese uso no está restringido. Pero yo no podría copiar la descripción (receta) y venderla o regalarla, porque justamente eso está prohibido. Esta forma de protección está pensada más que nada para obras literarias y audiovisuales.

Finalmente, en todo el mundo se ha decidido que el derecho de autor es el que también rige al software, con algunos temas específicos tales como derecho a copias de respaldo, a modificaciones para lograr interoperabilidad, etc. En EE.UU. y algunos otros países se ha optado por proteger también el software por patentes, pero ese es un tema aún no resuelto ni menos consensuado.

¿En qué se diferencia la noción de propiedad intelectual de un producto farmacéutico, de un libro, de un software?

En un producto farmacéutico, no tiene sentido aplicar el derecho de autor, porque sin copiar ni la forma de la pastilla (u otro producto) ni la receta, es muy fácil hacer una pastilla con idéntico efecto y diferente forma, u obtener la receta a partir de una pastilla legítimamente comprada. Lo que se quiere hacer en este caso es que toda la millonaria inversión en investigación previa, que resulta en la receta casi trivial de cuánto mezclar de cada componente, pueda cobrarse en el precio de las pastillas. De esa forma, los laboratorios tienen incentivos para arriesgar muchos recursos en investigaciones, para que alguna de ellas resulte en ingresos futuros suficientemente atractivos como para compensar los fracasos y la incertidumbre.

En el caso de un libro y de un software, lo que importa es que no se pueda llegar y copiar, sino que el creador original tenga el derecho exclusivo de generar o permitir copias. Al contrario de lo que ocurre en el caso del producto farmacéutico, cualquiera puede (salvo que existan además otros mecanismos legales en juego) mirar el comportamiento del producto y crear el suyo propio, siempre y cuando no lo copie textualmente. Claro que en esto existe un área gris, dado que no es fácil replicar algo que ya he visto asegurando que no estoy "copiando" nada de la expresión original, sobre todo cuando los lenguajes de programación exigen que ciertos procesos se describan de forma idéntica o al menos muy similar.

¿Esta discusión es puramente "global" o tiene también características particulares en cada país y región? si es así, ¿cuáles son esas en nuestro país? ¿en Latinoamérica?

La legislación por regla general aplica dentro del área geográfica que determina el país o incluso la región dentro de un mismo país. Pero en el caso de los bienes intangibles, desde hace mucho tiempo se han generado tratados internacionales donde se apunta a unificar los criterios aplicados.

Inicialmente nuestra legislación era más cercana a la europea, por eso se llama de "derecho de autor", que se origina en la revolución francesa, y no "derecho de copia" o copyright que viene de la tradición anglosajona (EEUU, Inglaterra y sus colonias). Pero hoy en día las diferencias son cada vez más sutiles.

¿Nos podrías contar cuáles son las políticas del Gobierno chileno respecto de estos temas? ¿Hay desarrollo propio, o como en muchas áreas, sólo tomar la mejor política desarrollada en otras latitudes?

En general no hay mucha libertad para desarrollar políticas propias. Tanto los tratados internacionales amplios (convención de Berna en particular) y los tratados de libre comercio bilaterales

incluyen mucha reglamentación, y es en estas instancias donde se puede hacer algo. Al respecto, Chile tiene una participación activa, pero tal como sucede en todo el mundo, hace mucha falta el involucramiento de una parte que, a pesar de ser la más importante, ha sido generalmente obviada: la sociedad civil.

El Movimiento del Software Libre

¿Nos podrías contar sobre la organización en Chile y en la región de la gente que está involucrada en este movimiento de software libre?

En el tema del desarrollo de software libre no existe una autoridad central que controle nada, y eso es justamente la gran característica del FLOSS (Free / Libre / Open Source Software), con sus pros y contras. En general existen organizaciones que se generan en base a los intereses de sus participantes, y tienen vida propia. No se va a encontrar un representante del FLOSS en Chile, ni en Latinoamérica, ni en ninguna parte, sino más bien grupos organizados tras un objetivo o interés específico. Esto incluye grupos de desarrolladores de plataformas o de software específico, usuarios y facilitadores de tecnologías, empresas que participan en este ecosistema, etc.

A nivel mundial hay dos grandes referentes: la Free Software Foundation, de Richard Stallman, y la Open Source Initiative. A nivel chileno existen algunos referentes como el Centro de Software Libre (csol.org), la asociación gremial MundoOS (mundoos.com), portales de noticias como softwarelibre.cl, grupos de interés como Educalibre (educalibre.cl), proyectos de desarrollo como Chileforge (chileforge.cl) y Gnome Chile (gnome.cl) y otras como LinuxChillan, GNUChile o CDSL. También hay espacios de encuentro y discusión, dentro de los cuáles se destaca el Encuentro Linux (encuentrolinux.cl).

¿Cómo te involucraste en este movimiento y cuáles son tus motivaciones personales en esto?

Muy temprano en mi carrera me interesó esto de que había cierto software que yo podía copiar sin problemas y otro en el cual me prometían las penas del infierno si lo hacía. Como nadie fue capaz de contestarme si era verdad que el software "copiable" era solamente para uso no comercial o si había limitaciones de otra especie, tuve que investigar por mi cuenta. Hoy en día hay bastante más información al respecto y tengo clara la película, pero me interesa el FLOSS y otros movimientos semejantes (Open Access Journals, Creative Commons, Open Access Courseware) como fenómeno social y productivo. Por eso mismo actualmente estoy más relacionado con las políticas relativas a tecnologías de información que el software propiamente tal.

¿Cómo puedo participar como desarrollador, estudiante, profesor, interesado en el área, en el movimiento del software libre?

En el FLOSS impera la meritocracia (ejemplificada magistralmente en el dicho "show me the code"), así que lo mejor es meter las manos cuanto antes. Eso no implica necesariamente programar, hay muchas otras formas de participar: difundir software, ayudar a otros a instalarlo, reportar problemas y potenciales causas, participar en grupos de usuarios, etc. Existen muchas formas de participar, y muchos grupos de los cuales se puede formar parte (cosa que aporta en general más que formar un nuevo grupo separado del resto). Basta encontrar el que reúna los mayores intereses en común con los propios y participar en lo que más pueda aportar o que más acomode y se vea que hace falta. Las formalidades para ello, si las hay, varían de grupo en grupo, pero aparecen de forma bastante clara en las descripciones del mismo.

Entrevista a Cláudia Bauzer Medeiros sobre la Sociedad Brasileña de Computación

por Claudia Páez, periodista del DCC.

Durante su visita al DCC en 2007, la docente y ex Presidenta de la Sociedad Brasileña de Computación, Cláudia Bauzer Medeiros, conversó con nosotros sobre el organismo que dirigió hasta ese mismo año, el rol de los posgrados en su país, de las peculiaridades de la Ciencia de la Computación, del papel de las mujeres en éstas y de los desafíos que la investigación en el área tiene trazados para la presente década en Brasil.

¿Cuáles son las principales tareas de la Sociedad de Computación que preside?

La Sociedad tiene varias tareas importantes. La primera es educación. En el nivel de posgrado tenemos una red con todos los cursos de posgrado del país, de todos los coordinadores, quienes tienen encuentros anuales para decidir la conducción del posgrado en computación de Brasil. También tenemos un examen nacional; todos los estudiantes que quieren hacer el posgrado en computación tienen que hacer nuestro examen. En el nivel de grado hacemos cursos para coordinadores de Bachelor o Computer Engineering. En general hay cursos nacionales para coordinadores con 300 ó 400 alumnos y también tenemos cursos regionales. Tenemos una Comisión de Educación que es un grupo de profesores que mantiene un estudio constante del currículum de computación. Los más de mil cursos de grado en computación siguen las recomendaciones que da la Sociedad en términos de contenidos.

Desde hace ocho años organizamos las Olimpiadas de Informática. Van niños de la escuela primaria desde los 10 años de edad hasta los 17 ó 18. Los 60 mejores del país van a una Universidad, que actualmente es la mía, a un curso de una semana en teoría y técnicas de programación, y al final de la semana hay un examen y los cuatro mejores son enviados a las Olimpiadas Internacionales. Hay también lo que llamamos Maratón, para alumnos en las universidades, y que al final participan en la competencia internacional ACM Collegiate en los EUA.

Cláudia Medeiros es docente y directora del Laboratorio de Sistemas de Información del Instituto de Computación de la Universidade Estadual de Campinas (UNICAMP), Brasil. Visitó Chile en 2007 invitada por el DCC para integrar el Comité Examinador de una defensa de tesis de doctorado. Ocupó la presidencia de la Sociedad de Computación de su país entre 2003 y julio del año pasado. Esta es una de las sociedades en su tipo más grandes de Sudamérica; fundada en 1978, en la actualidad está integrada por cerca de siete mil socios entre estudiantes, profesionales, docentes e investigadores del área, además de los socios institucionales como universidades y empresas nacionales y transnacionales. Aquí reproducimos parte de aquella entrevista realizada en esa oportunidad.

En el área de investigación tenemos 20 comisiones especiales que son grupos de interés que organizan congresos todos los años. La Sociedad promueve y da soporte logístico por año a 35 congresos, con un total de asistentes de 15 a 17 mil personas. Y apoyamos otros congresos que el

año pasado (2006) fueron 77. Tenemos manuales de cómo se organiza un congreso, un sistema de entrega de papers centralizado, y muchas otras actividades de apoyo a la investigación.

En la parte política, como Sociedad somos invitados por el gobierno a participar en comisiones que deciden, por ejemplo, la planificación en investigación en determinados asuntos, y lo que yo creo que es muy importante, la parte de reglamentación de la profesión.

¿Cómo está organizada la Sociedad de Computación?

Tenemos un grupo de directores; son investigadores de universidades. Cada uno se encarga de una parte de lo que hace la Sociedad. Hay direcciones Administrativa, de Educación, de Congresos, de Secretarías Regionales, de Publicaciones, de Reglamentación, etc. Existe una Dirección que se encarga de las relaciones con el exterior, que busca incluso convenios con empresas, que se interesan por determinados congresos organizados por la Sociedad, entonces dan fondos para apoyar su organización. El gobierno también concede fondos para algunos congresos que organizamos.

¿Cuál es la ventaja de tener como a uno de sus socios a la empresa privada?

A través de esta asociación las empresas, primero, tienen acceso a nuestras publicaciones e investigadores. Tienen también un canal de comunicación con los socios que así lo autorizan. Muchas de ellas tienen fondos para invertir, entonces tienen interés en los descuentos de los impuestos o por la publicidad que se genera. Para la Sociedad, la ventaja es que las empresas también son parte de la sociedad brasileña; además, su participación como asociados permite a las empresas conocer mejor la vida académica y por lo tanto tener contactos más directos con los investigadores brasileños.

Los Cinco Desafíos de la Investigación en Computación

Por lo descrito, en Brasil se asume que no se puede avanzar en el campo tecnológico sin reforzar y priorizar la actividad científica.

En Brasil esto es asumido por todos los científicos. Pero no necesariamente por otras personas. Porque tenemos tantos problemas en Brasil, problemas sociales. Entonces muchas veces dicen "tenemos un problema de educación, tenemos que solucionar el sistema brasileño de educación, total". Lo que sentimos también muchas veces es que a cada cambio de política, de gobierno, se empieza todo de nuevo. Tenemos una buena base de investigadores, pero hay necesidad de pensar más en el futuro a largo plazo.

¿Hay puntos comunes entre problemas tan básicos como el que usted señala, con lo que ustedes desarrollan como investigadores?

Claro, claro. El año pasado (2006) la Sociedad hizo un encuentro de 30 investigadores para planificar los cinco grandes desafíos de la investigación en computación para Brasil en la década 2006-2016. Y utilizamos como modelo de organización el mismo utilizado en Estados Unidos por la NSF (National Science Foundation).

¿Y cuáles son esos cinco desafíos?

El primero es el manejo de grandes volúmenes de datos multimedia distribuidos. El segundo, modelamiento computacional de sistemas complejos: artificiales, naturales, socio-culturales y de interacción humana con la naturaleza. El tercero, el impacto del cambio de la tecnología de construcción de computadores basados en silicio a nuevos tipos de computadores, por ejemplo, computación cuántica, biológica. El cuarto es acceso participativo y universal al conocimiento para el ciudadano brasileño: cómo desarrollar tecnología, software, base de datos, sistemas que permitan

al ciudadano brasileño no sólo acercarlo a la información sino agregar lo que sabe él a la información, y a participar en el progreso del país. Y el quinto desafío es desarrollo de sistemas extensibles, durables, confiables y ubicuos. En inglés la palabra correspondiente es "dependable". Los expertos en esta reunión dijeron que dependable no es sólo confiable, hay que ser durable, extensible, testeable. La traducción en inglés del desafío es "dependable and ubiquitous". En portugués ocupa tres líneas de descripción. Entonces uno de nuestros investigadores dijo por qué no usamos "sistemas omnivalentes" Y yo dije: perfecto. Es una palabra que no existe en portugués, pero me encantó (se ríe) y la utilizo siempre.

En Chile, aproximadamente el 80 por ciento de la capacidad de investigación e innovación del país proviene de las universidades ¿Este porcentaje es similar en relación a las universidades de Brasil?

Creo que sí. El Google Research se estableció en Belo Horizonte y está activamente contratando a doctores recién egresados, en todo el país. Hay unos cuantos centros de investigación del gobierno que no están en universidades, por ejemplo, el INPE brasileño, que es el correspondiente de la NASA y que concentra la investigación aeroespacial. La EMBRAPA es un importante centro de investigación en agricultura. Hay otros tantos centros de investigación del gobierno, en varios niveles, donde hay doctores.

Me imagino que para que la Sociedad haya alcanzado el nivel en el que trabajan actualmente, con una red de investigadores y universidades nacional, organizando constantemente actividades, influyendo en la agenda legislativa, debieron haber realizado varias conquistas.

Yo creo que hay algunos marcos, varios presidentes consiguieron avances, y uno de los grandes fue cuando Flávio Wagner, el presidente a fines de los noventa, profesionalizó muchísimo la Sociedad. Pero también Brasil está creciendo mucho en términos de investigación y de alumnos, de cursos, que han tenido necesidad de más servicio y apoyo de nuestra Sociedad. Nuestro problema ahora es cómo poder organizar las cosas que están bajo la Sociedad pero de forma fluida. Yo creo que la presidencia de Wagner coincidió también con la época en que el gobierno creó bajo el Ministerio de Ciencia y Tecnología algo que existe hasta ahora pero no de la misma forma, llamada SEPIN: "Secretaría de Política de Informática". Es una especie de secretaría extraordinaria del gobierno que se encarga de todo lo que es tecnología de información. Muchos de los principales personajes de esta organización en la época de Flávio Wagner eran nuestros socios o reconocían la importancia de la Sociedad desde el punto de vista de la investigación y la enseñanza, nos llamaban y nos involucraban en muchas cosas como organismo. Y con eso conseguimos ocupar algunos espacios políticos. Aún nos faltan muchísimos, pero ya somos reconocidos en varios lugares.

¿Cuál son los desafíos que su país enfrenta en materia de Ciencias de la Computación?

Ya te hablé de los cinco desafíos, y se suma un desafío medio científico, medio político, que es el reconocimiento de la computación como un área que tiene sus características y que debe ser reconocida de forma especial. Según la clasificación oficial brasileña de áreas de investigación, definida por el Ministerio de Ciencia y Tecnología hace más de 30 años, la Computación está debajo de las Ciencias Exactas y de la Tierra, junto con las matemáticas, física. Pero todos los investigadores y todos los ministerios se dan cuenta que cada vez más aumenta la necesidad de computación para el progreso del país en términos tecnológicos y científicos. Los científicos de otras áreas nos buscan para que los ayudemos a desarrollar sus proyectos. Pero la mayoría de esas personas nos piden programas, instalación de software, servicios. Entonces hay un desafío doble: de nuestro lado,

los investigadores en computación precisan también aprender con los otros investigadores, porque no podemos hacer investigación con otras áreas sin estudiarlas. Hay que compartir y cooperar. Y el desafío para nuestros hermanos investigadores es reconocer la computación; darle el espacio que merece en la organización del país, en las áreas de investigación. El desafío político es ser reconocidos en nuestras peculiaridades.

Alianza Universidad-Empresa

¿Existen alianzas activas de cooperación y trabajo entre universidades y el sector privado en el área de la innovación? ¿En qué se traducen?

En varias universidades sí. Te puedo citar principalmente mi Universidad, Unicamp, donde fue creada la Agencia Inova. El objetivo es buscar convenios con empresas para atraer fondos a proyectos de investigación de la Universidad. Le muestran a las empresas lo que nosotros podemos hacer, y nos anuncian los intereses que tienen las empresas. La segunda cosa que Inova hace es ayudar a los investigadores de Unicamp a hacer pedido de patentes. El año pasado (2006) Unicamp fue la institución brasileña que más depositó patentes en Brasil, incluso en comparación con todas las empresas, con la industria del petróleo.

¿Incluye esto la participación del Departamento de Informática?

No. Primero porque en patente de software no tenemos cultura ¿qué es patentar un algoritmo? no sabemos. Hay algunos procesos en los que estamos involucrados, y a lo mejor podríamos obtener patentes. Pero no es parte de nuestra cultura. Tenemos patentes de producción de alimentos, fibra óptica y muchas otras cosas en la Universidad.

Algo que me pareció muy interesante es que la Agencia de Apoyo a la Investigación en el Estado de San Pablo (FAPESP), firmó un convenio con Microsoft Research, en los Estados Unidos, para crear un Instituto Virtual de Investigación en Tecnología de Información con un millón de dólares, y ya lanzaron el primer llamado de proyecto para financiar investigación en el estado. Esto es interesante porque es un modelo distinto al que tenemos en Brasil.

La segunda cosa de muchísimo interés es que el tópico de la llamada de proyectos es el cuarto desafío de la Sociedad Brasileña de Computación. Entonces este es el reconocimiento oficial de lo que hacemos en la Sociedad.

En Brasil existe también en varios grupos una especie de mentalidad que lo público no debe asociarse con lo privado. Claro que no debe hacerlo, pero hasta un cierto nivel. Pero cuando la asociación es solamente crear un instituto de investigación, en la que el privado es serio y trae plata nueva para investigación para el país, en el país, no veo por qué no aprovechar la oportunidad.

En Chile se creó el Consejo de Innovación, que entre sus objetivos está proponer la forma y las materias en que se invertirán los impuestos recaudados por el Estado provenientes de un royalty aplicado a las empresas mineras. ¿Brasil cuenta con fondos permanentes estatales destinados a cubrir investigación?

Sí. Yo no sé exactamente los números pero tenemos por ejemplo el Fondo del Petróleo para Investigación e Innovación en Petróleo. El Fondo de Agricultura, de Biotecnología, y el Fondo de Informática también, etc. El de Informática es sacado de las empresas de Informática, que pueden descontar la plata de sus impuestos.

Los Estudiantes, de Hoy y Ayer

¿Cuál es la importancia de los posgrados en Brasil, considerando que en Computación al año cien personas obtienen su doctorado?

Es importante porque cada vez más la industria quiere gente con maestría. Para los doctores el mercado de trabajo son las universidades, pero esperamos que poco a poco las empresas empiecen a pedir doctores para actividades de más largo plazo. El número de cargos es limitado, por eso hay que encontrar otras oportunidades de trabajo para ellos.

¿Las tesis de doctorado en su universidad son netamente teóricas, o tienen relación con problemas prácticos?

Esto varía muchísimo. Las hay totalmente teóricas y hay tesis que son muy aplicadas, por ejemplo, involucrando problemas de criptografía o de solución de problemas en bases de datos. Incluso tenemos becas en mi Universidad dadas por empresas.

¿A condición de que el estudiante haga una tesis referente a ciertos problemas de la empresa?

A veces sí, a veces no. Pero, en el caso de la maestría, muchas empresas vienen y están dispuestas a pagar becas de dos años para determinados problemas. Si hay profesores que se interesan por esos problemas, luego van a la empresa a evaluar y dicen "sí, yo estoy interesado en este problema" y recibimos el fondo para la beca. Para el doctorado es un poco más difícil por el tiempo de duración.

¿La actitud de los estudiantes de hoy para enfrentar la Ciencia de la Computación es la misma que tenían los estudiantes hace 10 años?

No. El estudiante de hoy está muchísimo más habituado a las herramientas y a la web; todo muy fácil y listo para usar. Si le damos un programa para hacer, no tiene el hábito de planificar cómo lo va a programar; comienza directamente e intenta 50, 60, 70 veces, y lo puede hacer en un día. Mientras que yo cuando empecé a programar tenía el resultado al día siguiente, no podía esperar 200 días para ejecutar 200 veces un programa. El estudiante de hoy sí, y esta actitud naturalmente tiene ventajas porque facilita la prueba y la participación del usuario, pero no le enseña a planificar y esto no es bueno. No tienen paciencia. Otra cosa buena es que hoy el estudiante está más habituado a contestar, a tener actitudes independientes. Hoy nos preguntan por qué. Son más críticos.

Ya finalizando la entrevista, Cláudia Bauzer contó que una de las problemáticas que ella abordó a la cabeza de la Sociedad de Computación es la decreciente inclusión de mujeres en esta área. Según la investigadora, "se está haciendo un esfuerzo internacional para atraer mujeres al trabajo en computación, porque esto es un fenómeno mundial. Una primera cosa es convencerlas de que el trabajo en computación no es sólo quedarse frente a un computador para programar. Las empresas dicen que las mujeres en general son mejores para trabajar en contacto con los clientes, son más pacientes. Hay que hacer un estudio para investigar cuáles son los factores que están influyendo. Ahora es prioridad de varias empresas de Estados Unidos e Inglaterra, y también en Brasil, la contratación de mujeres en informática. En Estados Unidos están comenzando a trabajar con niñas de primaria; contándoles cómo es trabajar en computación. Nosotros vemos esto en las Olimpiadas de Brasil, porque hasta los 14 años de los finalistas de todo el país, que van a mi Universidad a hacer los cursos, la mitad son mujeres. A partir de los 14, ellas son el diez por ciento y a partir de los 16 hay cero mujeres. Entonces qué pasó con las niñas, que eran el 50 por ciento de los mejores en Brasil? ¿qué pasó?

Grupo MaTE de Ingeniería de Software

Desde hace algunos años tiene lugar en nuestro Departamento una sostenida actividad en el área de Ingeniería de Software. La coincidencia de investigadores y estudiantes de posgrado en torno al modelamiento de software y temas asociados provocó una intensificación en la investigación y trabajos conjuntos en el área. Esta situación fue recientemente formalizada a través de la creación del grupo MaTE.

MaTE nació oficialmente en enero de 2008, y sus principales materias de estudio son el Desarrollo de Software Dirigido por Modelos (MDD), Transformaciones de Modelos (MT), Líneas de Productos de Software (SPL), Líneas de Procesos de Software (SPrL), Arquitecturas de Software (SA) y Modelamiento de Atributos de Calidad. Su nombre es acrónimo de Model and Transformation Engineering, y es un denominador común del interés de sus integrantes por diversas áreas de la Ingeniería de Software: modelamiento de software aplicado a diferentes dominios y ámbitos, y procesamiento sistemático de modelos producidos mediante transformaciones de modelos.

Áreas de Investigación

En ese contexto, el objetivo de MaTE es contribuir al avance y maduración de la Ingeniería Dirigida por Modelos (MDE), área emergente dentro de la Ingeniería de Software. Actualmente el Grupo busca alcanzar dicho objetivo por dos vías separadas y a la vez complementarias. Por un lado se encuentra la aplicación de técnicas existentes a diferentes aspectos del desarrollo de software, como la definición y gestión de líneas de productos, la instanciación de modelos de procesos de desarrollo y la definición de arquitecturas de software. Estas actividades presentan un marco dual en el que se incorporan técnicas nuevas a los ámbitos mencionados y se descubren desafíos propios de la puesta en práctica de MDE. Por otro lado, el estudio de conceptos básicos tales como la noción de modelo, su definición y sus transformaciones, permite refinar las ideas fundamentales y dar lugar a la propuesta de mejores prácticas tendientes a atacar los desafíos identificados. MaTE busca ser referente en su actividad a nivel regional, y un mecanismo para ello es comenzar a tener presencia en las principales conferencias internacionales en el área de MDE.

Integrantes

El Grupo MaTE es numeroso y heterogéneo. Actualmente está integrado por dos profesoras de jornada completa, Cecilia Bastarrica (Uruguay) y Nancy Hitschfeld (Chile); cuatro estudiantes de

doctorado: Pedro Rossel (Chile), Julio Hurtado (Colombia), y Andrés Vignaga y Daniel Perovich (Uruguay); cinco estudiantes de magíster: Andrés Astaburuaga, Claudio González, Marco Ribo, y Sebastián Rivas (Chile), y Jesica Madrid (Ecuador); y un estudiante de pregrado: Cristian Rojas (Chile). Asimismo cuenta con la colaboración como investigadora asociada de Jocelyn Simmonds (Chile), quien actualmente estudia un doctorado en la Universidad de Toronto, Canadá.

Logros

A los pocos meses de su formalización, MaTE ya cuenta con publicaciones en diferentes conferencias de primer nivel en el área de MDE, tales como el workshop de Calidad en Modelamiento y el Simposio Doctoral, ambos asociados a MoDELS'2007; el track de Transformaciones de Modelos del Simposio de Computación Aplicada (SAC'2007), y su nuevo formato, la Conferencia Internacional de Transformaciones de Modelos (ICMT'2008); además de la Conferencia Internacional de Reutilización de Software (ICSR'2006). Asimismo, el Grupo ha tenido presencia en revistas internacionales de Ingeniería de Software como International Journal in Software Engineering and Knowledge Engineering, y Advances in Engineering Software.

MaTE cuenta con numerosos contactos con investigadores de centros de investigación internacionales de primer nivel en MDE, entre los que se destacan la Universidad de Waterloo (Canadá), Universidad de Rennes (Francia), Universidad de L'Aquila (Italia), y Universitat Oberta de Catalunya (España).

Actualmente el Grupo participa en el Proyecto Tutelkán financiado por CORFO (Corporación de Fomento de la Producción del gobierno de Chile), conjuntamente con la Universidad Federico Santa María. A través del proyecto se busca elaborar un modelo de procesos de desarrollo de software para las PyMEs de software en Chile. Asimismo, MaTE se encuentra elaborando un proyecto que involucra a la Universidad de Chile, Universidad de la República (Uruguay) y el laboratorio INRIA (Francia).

Información y contacto http://mate.dcc.uchile.cl/

PLEIAD: Explorando Nuevos Lenguajes para Mejores Programas

PLEIAD (Programming Languages and Environments for Intelligent, Adaptable and Distributed Systems) está dedicado a explorar en qué forma los lenguajes de programación y sus ambientes de desarrollo pueden permitir la construcción de software evolucionable y adaptable. Considerando fundamentalmente contextos desafiantes como la computación distribuida y ubicua.

Los lenguajes de programación a medida que son utilizados van adquiriendo cierta popularidad. Hoy Java y C# son muy populares. Su antecesor en popularidad fue C++ y los antecesores a este fueron C y FORTRAN sucesivamente. Los lenguajes de programación son diseñados para satisfacer determinados ambientes tecnológicos de ejecución, con el propósito de facilitar el desarrollo de ciertos tipos de soluciones. A medida que la tecnología va cambiando, los lenguajes de programación van quedando obsoletos. Es por esto que un lenguaje de programación puede verse afectado por cambios y evoluciones a lo largo de su vida. La comunidad de Ingeniería en Software y Lenguajes está constantemente en búsqueda de lenguajes de programación que satisfagan las nuevas tecnologías. Por ejemplo, ahora que Java ya tiene más de 10 años, se está explorando lo que a veces se denomina el "mundo post-Java". Otro ejemplo claro de esta búsqueda es la utilización de forma más masiva de los lenguajes de programación dinámicos como Python y Ruby.

Es importante señalar que los lenguajes de programación tienen un rol fundamental en la forma en la cual los desarrolladores manejan la complejidad de los sistemas computacionales. Por ejemplo, la programación por objetos y por componentes es la base sobre la cual muchos sistemas de gran envergadura están construidos. La ubicuidad incesante de la computación a todos los niveles de la sociedad implica más complejidad para el software. Esto llama a desarrollar mejores formas lingüísticas que manejen esa complejidad. Un ejemplo de este fenómeno es la emergencia del paradigma de Programación Orientada a Aspectos (AOP), el cual permite la definición modular de preocupaciones que son transversales a los objetos de un sistema, como la seguridad o la coordinación de actividades. Con AOP, se logra mejor desacoplamiento y reusabilidad de los componentes y objetos. Esto también permite mejor adaptabilidad del sistema en forma dinámica.

Principales Áreas de Investigación

En el laboratorio PLEIAD se investigan al día de hoy tres áreas principales:

- Programación por aspectos (AOP)
- Depuración y comprensión de programas

- Programación de sistemas de computación pervasiva.

A continuación detallamos brevemente estas áreas.

Programación por Aspectos (AOP). AOP introduce nuevos mecanismos para definir software en forma más modular y más adaptable. AOP tiene una fuerte herencia de los trabajos sobre reflexión computacional, uno de los temas sobre los cuales integrantes de PLEIAD han trabajado previamente. En el contexto de AOP, PLEIAD participa de la investigación sobre definición de lenguajes AOP. En efecto, aunque la comunidad ya cuenta con un lenguaje AOP de alcance industrial llamado AspectJ, muchos temas quedan por explorar para revelar toda la potencialidad de los aspectos en el desarrollo de software. Actualmente dos temáticas principales se desarrollan en PLEIAD: la construcción y definición de lenguajes AOP, incluyendo lenguajes específicos a cierto dominio (por ejemplo un lenguaje dedicado a la definición del manejo de transacciones), y el mejoramiento de los mecanismos provistos para el control del impacto de un aspecto sobre un sistema dado. Esta investigación se hace con una mirada particular a las problemáticas asociadas a la introducción de aspectos en sistemas complejos, es decir, concurrentes y distribuidos.

Depuración y comprensión de programas. La depuración consiste en ayudar al desarrollador a encontrar errores en un programa. De forma más general, esto pertenece al área de "entendimiento de programas" (program understanding). Es decir, cómo ayudar a un ser humano a entender con suficientes detalles lo que pasa en un programa como para ser capaz de modificarlo o corregirlo, manteniendo un nivel de abstracción que le permita aprender el programa sin ser sobrepasado por su complejidad. En particular, en PLEIAD se está trabajando sobre un sistema de Depuración Omnisciente llamado TOD (Trace-Oriented Debugger). Es decir un depurador donde se puede navegar en la historia de ejecución de un programa, tanto hacia adelante como atrás en el tiempo. Esto permite recorrer vínculos causales que son muy difíciles de reconstituir en los depuradores actuales. Mediante la utilización de un depurador omnisciente podemos saber en qué momento y en qué contexto, determinada variable fue asignada un valor inválido, lo que con un depurador normal no se puede saber.

Computación Pervasiva. La computación pervasiva o "inteligencia ambiental", se refiere al desarrollo de sistemas computacionales para usuarios móviles con aparatos móviles, en los cuales la integración de los sistemas en la vida diaria es lo más transparente posible. Esto requiere de sistemas que sepan captar su ambiente de ejecución y adaptarse a ello en forma dinámica. Programar dichos sistemas con lenguajes tradicionales implica un nivel de complejidad enorme para manejar todos los detalles relacionados con la naturaleza volátil del ambiente y de las conexiones. PLEIAD trabaja en lenguajes dedicados, que proveen abstracciones adecuadas para que el programador pueda especificar tanto la percepción del ambiente cómo la adaptación del sistema. En estos momentos se experimenta con el lenguaje AmbientTalk, desarrollado en el laboratorio PROG de la Vrije Universiteit Brussel (Bélgica) con el cual PLEIAD está colaborando. También se exploran técnicas de inteligencia artificial aplicadas a computación pervasiva.

Historia, Logros y Desafíos

El laboratorio PLEIAD fue inaugurado recién en noviembre de 2007, con la presencia de Ron Goldman, Ingeniero de Investigación de Sun Labs en EE.UU., a cargo del proyecto de Sun SPOTs; unos aparatos para inteligencia ambiental dotados de sensores y que corren una máquina virtual Java.

PLEIAD está inicialmente formado por dos profesores full-time, Johan Fabry y Éric Tanter, cuatro estudiantes de doctorado y un estudiante de magister. Y se están integraron nuevos miembros, principalmente provenientes de posgrado.

A pesar de ser un laboratorio joven, PLEIAD ya está logrando exponerse a nivel internacional en las distintas áreas en que se desempeña. En la conferencia ACM sobre desarrollo de software orientado a aspectos, organizada en abril (AOSD 2008), se presentó un artículo técnico sobre "scoping" de aspectos dinámicos, se organizó un workshop sobre lenguajes de aspectos específicos a dominios, y se hizó una demostración del Depurador TOD. Un artículo técnico sobre el depurador omnisciente ha sido publicado en OOPSLA 2007, la conferencia de referencia del ACM en el área, y otro, sobre uso de TOD para programas orientados a aspectos, se presentó en el simposio de computación aplicada del ACM en marzo (SAC 2008). También se organizó un workshop sobre computación inspirada en la biología en diciembre en Valparaíso (BIC 2007), Chile, que contó con la presencia de varios expertos internacionales, tanto de Europa como de EE.UU. PLEIAD está involucrado como miembro fundador de la red de colaboración sudamericana Latin AOSD, y participa en otros proyectos de cooperación internacional, en particular con el INRIA de Francia.

Información y contacto en http://pleiad.dcc.uchile.cl/

Instituto Virtual LACCIR: una Red de Investigación para Latinoamérica y el Caribe

El Instituto Virtual LACCIR (Latin American and Caribbean Collaborative) se creó en mayo de 2007 gracias a un acuerdo firmado por las facultades de Ciencias Físicas y Matemáticas de la Universidad de Chile, de Ingeniería de la Pontificia Universidad Católica de Chile, y Microsoft Research.

Uno de los objetivos principales del Instituto es establecer una red donde el trabajo de los académicos trascienda las fronteras de la universidad, de modo de llevar a cabo proyectos de investigación entre académicos de diversos países de Latinoamérica, compartir experiencias de aprendizaje a través de Internet y estrechar vínculos de colaboración entre las entidades participantes.

Durante los tres primeros años la administración del Instituto Virtual estará alojada en la Universidad Católica y en los tres siguientes se ubicará en la Universidad de Chile, específicamente en nuestro Departamento. La coordinación con el resto de los planteles y países, y las principales decisiones que toma el Instituto están a cargo del comité estratégico de LACCIR, donde participan la Universidad de Chile, la Pontificia Universidad Católica de Chile, Microsoft Research, el Banco Interamericano de Desarrollo y la Organización de Estado Americanos (OEA).

La Dirección Ejecutiva del Instituto está a cargo del académico de la Universidad Católica profesor Ignacio Casas, mientras que Claudia Leiva, de la misma universidad, ejerce como gerente general. En el comité directivo participan también los académicos del DCC de la Universidad de Chile Sergio Ochoa y José Pino, y el profesor de la Universidad Católica Yadrán Eterovich.

Funcionamiento

LACCIR funciona mediante una red instalada sobre Internet I e Internet II. Cada universidad participante tiene un nodo de videoconferencia conectado a dicha red virtual, lo que permite llevar a cabo la colaboración entre universidades. En el caso de la Universidad de Chile, el nodo está alojado en el Departamento de Ciencias de la Computación. El modelo de infraestructura que ocupa dicha red es el de "hub and spokes". En el "hub" se coordinan las principales tareas del Instituto, mientras los "spokes" constituyen los colaboradores distribuidos. Esta estrategia de organización busca acrecentar las relaciones entre universidades ubicadas en centros neurálgicos de investigación -que además cuentan con recursos para ello-, con aquellas con menores posibilidades de efectuar este tipo de proyectos. El Instituto Virtual cuenta con un fondo inicial de 935 mil dólares -aportados por Microsoft- de los cuales 180 mil se destinaron a implementar la infraestructura de colaboración necesaria para su funcionamiento.

Primera convocatoria de proyectos

Uno de los grandes objetivos de LACCIR es promover y otorgar fondos para el desarrollo de proyectos en Latinoamérica y el Caribe. Dichos proyectos deben estar enfocados a investigar el uso de soluciones tecnológicas para resolver problemas económicos y sociales comunes a la región.

Siguiendo esta línea, en octubre de 2007 LACCIR realizó la primera convocatoria para presentar proyectos siendo el requisito principal para la aceptación de las propuestas, que cada una contara con un equipo de investigadores de al menos dos universidades de distintos países de Latinoamérica y el Caribe. De esa manera se buscó fomentar la colaboración entre los académicos de la región.

En marzo de este año el Instituto Virtual dio a conocer los resultados de los proyectos favorecidos -en total cinco-, todos enfocados en el uso de las tecnologías para resolver problemáticas en las áreas de educación, salud, E-Government, agro-industria y cadenas productivas. En esta primera convocatoria, tres de los proyectos favorecidos son de coautoría de los académicos del DCC: José Pino, Luis Guerrero y Sergio Ochoa. Dichos proyectos son de un año de duración.

El fondo destinado al desarrollo de estos proyectos es de 50 mil dólares por proyecto, y un año de duración. El financiamiento inicial para proyectos lo provee Microsoft Research y está comprometido por los próximos dos años. Sin embargo, está la necesidad de acrecentar estos fondos y perpetuarlos para asegurar la trascendencia de LACCIR, lo que será posible en la medida que los proyectos desarrollados con fondos del Instituto sean exitosos. Además se espera que tengan un impacto positivo sobre Latinoamérica, ya que esto debiera permitir atraer a instituciones que deseen inyectar fondos a la investigación científica en computación aplicada.

Proyectos Ganadores, primera convocatoria

Una Herramienta Digital para Apoyar la Colaboración Asincrónica

Universidad de Chile; Universidad del Cauca, Colombia.

Asumiendo que cada día más estudiantes utilizan Notebooks o TabletPCs, con este proyecto se intentará desarrollar una herramienta "workbook" digital ejecutable en un TabletPC que les permita a los estudiantes compartir y sincronizar sus anotaciones digitales con sus compañeros e instructores, lo que les permitirá trabajar de forma colaborativa y por tanto mejorar su aprendizaje. Este workbook digital usará los servicios de OneNote 2007 y extenderá la plataforma Conference XP para integrar el trabajo asincrónico en ésta. La herramienta a desarrollar trabajará de dos formas: conectada a una sesión de Conference XP, y/o conectada a otra TablePC (punto-a-punto).

Jardin: Just an Assistant foR instructional DesIgN (Asistente para Diseño Instruccional)

Universidad de la República, Uruguay; Universidade Federal do Rio Grande do Sul, Brasil; Universidade Estadual de Londrina, Brasil; Universidad Nacional de Rosario, Argentina; Escuela Superior Politécnica del Litoral ESPOL, Ecuador, y Universidad Nacional Autónoma de México.

Este proyecto está enfocado al desarrollo de una herramienta que facilite la creación, descripción, búsqueda y reutilización de Objetos de Aprendizaje por parte del instructor, ya que a pesar de que la mayoría de los investigadores concuerdan en el valor de las Tecnologías de Objetos de Aprendizaje, la falta de madurez en las herramientas para usuarios finales se refleja en el bajo nivel de adopción de estas tecnologías entre instructores y aprendices.

Sistemas de Visión por Computador de Bajo Costo para la Evaluación de Calidad de Productos Alimenticios en Pequeñas y Medianas Empresas

Universidad Católica de Chile; Instituto Politécnico Nacional de México.

La industria de alimentos es uno de los sectores económicos más relevantes en Latinoamérica, que en la actualidad debe responder a las crecientes demandas de los consumidores por más información sobre la calidad y seguridad de los productos. Para cumplir este requerimiento, los sistemas de visión por computador cumplen un rol clave en cuanto permiten entregar una medición objetiva de atributos visuales relevantes como, por ejemplo, forma y color. A través de este proyecto, se busca potenciar el desarrollo de pequeñas y medianas empresas productoras de alimentos, mediante del diseño de sistemas de visión por computador de bajo costo. Estos sistemas aprovechan la infraestructura base de equipos comunes ya utilizados por la industria como escáneres y cámaras digitales, de modo de perfeccionar las herramientas de evaluación de calidad y seguridad de los productos alimenticios.

Colaboración en los Esfuerzos de Socorro en caso de Desastres

Universidad de Chile; Universidade Federal do Rio de Janeiro, Brasil.

Este proyecto busca entregar una solución efectiva y eficiente a quienes trabajan en labores de socorro después de ocurrido un desastre natural o causado por el hombre, como por ejemplo: terremotos, huracanes e inundaciones, entre otros. La solución propuesta implica el uso de Tecnologías de la Información, en particular computación móvil, sistemas de colaboración y redes móviles ad-hoc, para coordinar los distintos sistemas de información que poseen organizaciones como bomberos, policía, Cruz Roja y ejército, entre otras. De esa manera se busca agilizar los tiempos de respuestas y facilitar el trabajo en conjunto.

Teléfonos Inteligentes y Grandes Displays como Facilitadores de la Colaboración en el Trabajo en Hospitales

Universidad de Chile; CICESE Research, México.

El trabajo en hospitales se caracteriza por la necesaria coordinación y colaboración entre las diferentes áreas de especialización, el intenso intercambio de información, la integración de datos provenientes de múltiples dispositivos o equipos, y la movilidad de personal del hospital, así como también la de los pacientes, documentos y equipo. Con este proyecto se busca diseñar e implementar un ambiente de colaboración que soporte la interacción entre diversos dispositivos electrónicos, particularmente teléfonos inteligentes y grandes displays (ambient displays). Además de desarrollar un conjunto de servicios de software que permitan aumentar la colaboración informal co-localizada entre trabajadores nómadas en los hospitales. La idea es mejorar la coordinación y el proceso de toma de decisiones que llevan a cabo estos trabajadores nómadas -principalmente médicos y enfermeras- durante el proceso de atención a pacientes.

Programa de Doctorado en Ciencias mención Computación

Como Coordinación de Posgrado de nuestro Departamento nos complace presentarles el Programa de Doctorado en Ciencias mención Computación. Nuestra Universidad es una de las más prestigiosas instituciones de educación superior en Latinoamérica y el país. El 61 % de los presidentes de la República de Chile egresaron de nuestra institución; nuestros dos premios Nobel estudiaron aquí; y el 83 % de los Premios Nacionales en sus diversas disciplinas han laborado o estudiado en la Universidad de Chile.

Estamos ubicados en Santiago, la capital de Chile; ciudad concebida como un atractivo centro financiero con presencia de compañías nacionales e internacionales; un activo tráfico aéreo internacional y nacional, a pocas horas de vuelo del desierto, los glaciares o Rapa Nui, y a una hora de camino a la montaña y al mar; con una rica vida gastronómica y cultural. Asimismo Santiago se jacta de poseer un avanzado desarrollo en telecomunicaciones, y uno de los mejores índices de seguridad de la región.

El DCC es uno de los departamentos con mayor tradición en esta disciplina en Chile. Realizamos investigación del más alto nivel, y desarrollamos proyectos y productos informáticos utilizados por diversos sectores productivos nacionales. Todo en un ambiente creativo y entusiasta donde profesores y estudiantes trabajan estrechamente. Hoy nuestro Departamento congrega una veintena de académicos e investigadores de jornada completa, con magísteres y doctorados la mayoría obtenidos en Norteamérica y Europa. Ellos realizan investigación en áreas como: Algoritmos Geométricos; Bases de Datos; Criptografía y Seguridad; Estructuras de Datos; Informática Educativa; Ingeniería de Software; Interacción Humano-Computador; Investigación de la Web; Lógica y Aspectos Formales; Métodos Numéricos y Aplicaciones; Minería de Datos; Programación de Lenguajes; Recuperación de Información Multimedia; Redes, Sistemas Distribuidos y Paralelismo; Sistemas Colaborativos, etc.

El administrador nacional de dominio .Cl, NIC Chile; el centro de informática educativa, C5; el Grupo de Respuesta a Incidentes de Seguridad Computacional, CLCERT; el Centro de Investigación de la Web, CIW; el Instituto Virtual de Colaboración en Investigación LACCIR; el laboratorio Yahoo! Research Latin America; son ejemplos de nuestro prolífico quehacer científico.

Nos ubicamos entre los primeros departamentos de Chile en ofrecer programas de Magíster y Doctorado en Ciencia de la Computación; este último con 24 estudiantes activos en 2007, provenientes tanto del interior del país como de Argentina, Bélgica, Francia, México, Perú y Uruguay, entre otros países. Nuestros programas de posgrado cuentan con la acreditación oficial otorgada en Chile; certificación que avala a nuestros alumnos como postulantes elegibles para las becas otorgadas por el Estado, y otras importantes fuentes de cooperación internacional que financian el arancel y los gastos hasta en 100 por ciento. Nuestros estudiantes de Doctorado pueden solicitar varias becas internas que cubren el arancel y entregan fondos de apoyo que pueden llegar a cubrir completamente el costo de vida del alumno.

Extendemos la invitación para que nos visiten en: http://www.dcc.uchile.cl. Y de requerir mayor información, pueden solicitarla en: cpostg@dcc.uchile.cl

Coordinación de Posgrado DCC

Tesis de Doctorado Recientes

Métodos de Acceso y Procesamiento de Consultas Espacio-Temporales

Alumno: Gilberto Gutiérrez.

Profesores Guía: Gonzalo Navarro (Universidad de Chile), Andrea Rodríguez (Universidad de Concepción).

Fecha de Examen: Abril de 2007.

Existe una necesidad creciente por contar con aplicaciones espacio-temporales que necesitan modelar la naturaleza dinámica de los objetos espaciales. Las bases de datos espacio-temporales intentan proporcionar facilidades que permitan apoyar la implementación de este tipo de aplicaciones. Una de estas facilidades corresponde a los métodos de acceso, que tienen por objetivo construir índices para permitir el procesamiento eficiente de las consultas espacio-temporales.

Gilberto Gutiérrez en la actualidad se desempeña como director y profesor del Departamento de Ciencias de la Computación y Tecnologías de la Información, Facultad de Ciencias Empresariales, Universidad del Bío-Bío, Chile

En esta tesis se describen nuevos métodos de acceso basados en un enfoque que combina dos visiones para modelar información espacio-temporal: snapshots y eventos. Los snapshots se implementan por medio de un índice espacial y los eventos que ocurren entre snapshots consecutivos, se registran en una bitácora. Se estudió el comportamiento de nuestro enfoque considerando diferentes granularidades del espacio. Nuestro primer método de acceso espacio-temporal (SEST-Index) se obtuvo teniendo en cuenta el espacio completo y el segundo (SESTL) considerando las divisiones más finas del espacio producidas por el índice espacial. En esta tesis se realizaron varios estudios comparativos entre nuestros métodos de acceso y otros métodos propuestos en la literatura (HR-tree y MVR-tree) para evaluar las consultas espacio-temporales tradicionales (time-slice y time-interval). Los estudios muestran la superioridad de nuestras estructuras de datos en términos de almacenamiento y eficiencia para procesar tales consultas en un amplio rango de situaciones. Para nuestros dos métodos de acceso se definieron modelos de costos que permiten estimar tanto el almacenamiento como el tiempo de las consultas. Estos modelos se validaron experimentalmente presentando una buena capacidad de estimación.

Basándonos en nuestros métodos propusimos algoritmos para procesar otros tipos de consultas espacio-temporales, más allá de time-slice y time-interval. Específicamente diseñamos algoritmos para evaluar la operación de reunión espacio-temporal, consultas sobre eventos y sobre patrones

espacio-temporales. Se realizaron varios experimentos con el propósito de comparar el desempeño de nuestros métodos frente a otros propuestos en la literatura (3D R-tree, MVR-tree, HR-tree y CellList) para procesar estos tipos de consultas. Los resultados muestran un rendimiento, en general, favorable a nuestros métodos. En resumen, nuestros métodos son los primeros que resuelven de manera eficiente no sólo las consultas de tipo time-slice y time-interval, sino también varias otras de interés en aplicaciones espacio-temporales.

Indexación Efectiva de Espacios Métricos usando Permutaciones

Alumno: Karina Figueroa.

Profesores Guía: Gonzalo Navarro (Universidad de Chile), Edgar Chávez (Universidad Michoacana de San Nicolás de Hidalgo Mexico).

Fecha de Examen: Junio de 2007.

En muchas aplicaciones multimedia y de reconocimiento de patrones es necesario hacer consultas por proximidad a grandes bases de datos modelándolas como un espacio métrico, donde los elementos son los objetos de la base de datos y la proximidad se mide usando una distancia, generalmente costosa de calcular. El objetivo de un índice es preprocesar la base de datos para responder consultas haciendo el menor número de evaluaciones de distancia.

Los índices métricos existentes hacen uso de la desigualdad triangular para responder consultas de proximidad, ya sea partiendo el espacio en regiones compactas o utilizando distancias precalculadas a un conjunto distinguido de elementos. En esta tesis presentamos una nueva manera de resolver el problema, representando los elementos como permutaciones. La permutación se obtiene eligiendo un conjunto de objetos, llamados permutantes, y considerando el orden relativo en el que se ven los permutantes desde cada elemento a indexar.

Nuestra contribución principal es el haber descubierto que la proximidad entre elementos se puede predecir con mucha precisión midiendo la distancia entre las permutaciones que representan esos elementos.

Karina Figueroa en la actualidad se desempeña como profesora e investigadora, y administradora de Sistemas del Centro de Cómputo, Facultad de Ciencias Físico Matemáticas, Universidad Michoacana (UMSNH), México.

Una aplicación directa de nuestra técnica deriva en un método probabilístico simple y eficiente: Se ordena la base de datos por proximidad de las permutaciones de los elementos a la permutación de la consulta, y se recorre en ese orden. De la comparación experimental de esta técnica contra el estado del arte, en diversos espacios reales y sintéticos, se concluye que las permutaciones son mucho mejores predictores de proximidad que las técnicas hasta ahora usadas, sobre todo en dimensiones altas. Generalmente basta revisar una pequeña fracción de la base de datos para tener un alto porcentaje de la respuesta correcta.

Otra aplicación menos directa de nuestra técnica consiste en modificar el algoritmo exacto AESA, que por 20 años ha sido el índice más eficiente, en términos de cálculos de distancia, para buscar en espacios métricos. Nuestra variante, iAESA, utiliza las permutaciones para determinar el siguiente candidato a compararse contra la consulta. Los resultados experimentales muestran que es posible mejorar el desempeño de AESA hasta en 35 %. Esta técnica es adaptable a otros algoritmos existentes.

Se aplicó nuestra técnica al problema de identificación de rostros en imágenes, y se lograron resultados hasta ahora no alcanzados por los típicos algoritmos vectoriales usados en estas aplicaciones. Asimismo, dado que nuestra técnica no aplica explícitamente la desigualdad triangular, la probamos en algunos espacios de similaridad no métrica, obteniendo un índice que permite la búsqueda por proximidad con resultados semejantes al caso de los espacios métricos.

Minería de Datos en Motores de Búsqueda

Alumno: Marcelo Mendoza.
Profesores Guía: Ricardo Baeza (Universidad de Chile).
Fecha de Examen: Junio de 2007.

La Web es un gran espacio de información donde muchos recursos como documentos, imágenes u otros contenidos multimediales pueden ser accesados. En este contexto, varias tecnologías de la información han sido desarrolladas para ayudar a los usuarios a satisfacer sus necesidades de búsqueda en la Web, y las más usadas de estas son los motores de búsqueda. Los motores de búsqueda permiten a los usuarios encontrar recursos formulando consultas y revisando una lista de respuestas.

Marcelo Mendoza en la actualidad se desempeña como director y profesor de la Carrera de Ingenieria Civil Informática, Universidad de Valparaíso, Chile.

Uno de los principales desafíos para la comunidad de la Web es diseñar motores de búsqueda que permitan a los usuarios encontrar recursos semánticamente conectados con sus consultas. El gran tamaño de la Web y la vaguedad de los términos más comúnmente usados en la formulación de consultas son los principales obstáculos para lograr este objetivo.

En esta tesis se exploraron las selecciones de los usuarios registradas en los logs de los motores de búsqueda para aprender cómo los usuarios buscan y también para diseñar algoritmos que permitieran mejorar la precisión de las respuestas recomendadas a los usuarios. Se comenzó explorando las propiedades de estos datos. Esta exploración nos permitió determinar la naturaleza dispersa de estos datos. Además se presentaron modelos que permitieron entender cómo los usuarios buscan en los motores de búsqueda.

Luego, se exploraron las selecciones de los usuarios para encontrar asociaciones útiles entre consultas registradas en los logs. Los esfuerzos se concentraron en el diseño de técnicas que permitieran a los usuarios encontrar mejores consultas que la consulta original. Como una aplicación, se diseñaron métodos de reformulación de consultas que ayudaran a los usuarios a encontrar términos más útiles mejorando la representación de sus necesidades.

Usando términos de documentos se construyeron representaciones vectoriales para consultas. Aplicando técnicas de clustering se pudieron determinar grupos de consultas similares. Usando estos grupos de consultas, se introdujeron métodos para recomendación de consultas y documentos que permitieron mejorar la precisión de las recomendaciones.

Finalmente, se diseñaron técnicas de clasificación de consultas que nos permitieron encontrar conceptos semánticamente relacionados con la consulta original. Para lograr esto, se clasificaron las consultas de los usuarios en directorios Web. Como una aplicación, se introdujeron métodos para la mantención automática de los directorios.

Improving Learning-Object Metadata Usage During Lesson Authoring

Alumno: Olivier Motelet.
Profesor Guía: Nelson Baloian; José A. Pino (Universidad de Chile).
Fecha de Examen: Octubre 2007.

Para lograr coherencia y flexibilidad en unidades de aprendizaje basadas en documentos multimedia, varios autores han recomendado estructurar los componentes de los cursos en grafos. En un grafo de curso, los recursos educacionales son encapsulados como objetos de aprendizaje (LO - Learning Objects) con sus respectivos metadatos (LOM - Learning-Object Metadata) y son interconectados con relaciones de varios tipos retóricos y/o semánticos. Los grafos de recursos son almacenados en repositorios en los cuales los metadatos sirven para facilitar su recuperación y reutilización. Sin embargo, tales sistemas se enfrentan con problemas serios en cuanto al uso de los LOMs: los metadatos son difíciles de instanciar y los autores de cursos generalmente no tienen estímulos para cumplir con esta tediosa tarea ya que ellos mismos no se benefician de los metadatos que generan.

Olivier Motelet se desempeña como responsable de Indentificación de Comercialización Proyecto Innovador para Desarrolladores, Empresa IDTM, Francia.

La generación automática de metadatos resuelve este problema. Sin embargo, este método se limita a ciertos metadatos excluyendo la mayor parte de los metadatos subjetivos tales como los metadatos educacionales. Esta limitación motivó el enfoque de esta tesis sobre una técnica complementaria: un método híbrido basado en la sinergia entre procesos automáticos e intervención

humana. La generación híbrida de LOMs puede ser aplicada sobre los atributos que no pueden ser automáticamente generados. Sin embargo, este enfoque está basado en la contribución de usuarios no siempre cooperativos, quienes necesitarían ver beneficios para motivar su participación.

Proponemos estudiar los usos de LOM durante la creación de cursos, no sólo desde la perspectiva de la generación híbrida sino también desde la perspectiva de los beneficios que pueden brindar los LOMs. Esta estrategia tiene como objetivo soportar una retroacción positiva en la cual los beneficios puedan motivar la generación de LOMs de buena calidad, y la buena calidad de los LOMs pueda mejorar los beneficios.

En particular, esta tesis investiga métodos para (1) integrar sin transición la generación híbrida de LOMs dentro de una herramienta de creación de cursos, (2) procesar un conjunto de LOMs aunque ciertos metadatos quedaran incompletos, incorrectos, o faltantes, (3) mejorar los resultados de los métodos clásicos de recuperación de LOs usando los metadatos de los LOs que componen un curso.

Desarrollamos una herramienta de código abierto para validar las propuestas de esta tesis. Experimentos preliminares mostraron que los LOMs pueden mejorar significativamente la recuperación de LOs adicionales durante el proceso de creación de cursos.

www.dcc.uchile.cl
dcc@dcc.uchile.cl

REVISTA
BITS DE CIENCIA
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN
Universidad de Chile Nº 1 2008