Gonzalo Navarro, primer Doctor en Ciencias de la Computación (al centro).

PRIMER DOCTORADO FORMADO EN EL DCC

El 3 de diciembre fue una fecha significativa para el Departamento de Ciencias de la Computación, debido a que ese día el alumno Gonzalo Navarro Badino, obtuvo el título de Doctor en Ciencias de la Computación, con nota 7 y distinción máxima otorgada por la Comisión de Examen de Grado que estuvo integrada por:

Profesor Ricardo Baeza-Yates:
Director de Tesis
Profesor Patricio Poblete:
Profesor Integrante
Profesor Jorge Olivos:
Profesor Integrante
Profesor Esko Ukkonen:
Profesor Invitado de la Universidad de Helsinki, Finlandia.

El profesor Gonzalo Navarro es el primer doctor formado en el DCC y la tesis presentada versa sobre «Búsqueda Aproximada en Texto» Un breve resumen del tema es el siguiente:

La cantidad de información textual electrónicamente accesible está experimentando un enorme crecimiento en los últimos años. El incremento en volumen y heterogeneidad del texto presente en las bases de datos textuales hace cada vez mas difícil dar garantías sobre su calidad. Si bien en algunos casos estos es posible (por ejemplo para una enciclopedia electrónica con estrictos controles de calidad), hay muchos casos donde el problema persiste (por ejemplo en el World Wide Web, una base de datos textual ad-hoc con un volumen cercano a los 60 gigabytes). Los errores cometidos al ingresar un texto, ya sea por simple digitación o por uso de software de OCR (reconocimiento óptico de caracteres), hacen que las palabras ingresadas erróneamente no puedan recuperarse más. Más aún, es posible que realmente se desee buscar con incerteza al no recordarse exactamente la forma en que se escribe un nombre.

La búsqueda aproximada en texto consiste en encontrar ciertos patrones en un texto (ambos vistos como secuencias de símbolos) permitiendo que el calce no sea exacto sino que contenga (una cantidad limitada de) errores.

Este es un problema central en muchas áreas, no sólo en recuperación de texto sino también en biología computacional (para comparar cuan parecidas son dos secuencias de ADN o proteínas), reconocimiento de patrones (por ejemplo en data mining) y en bases de datos multimedia (por ejemplo para buscar una señal de audio, donde el calce nunca es exacto).

El objetivo de esta tesis fue encontrar mejores algoritmos y estructuras de datos que los existente para las diversas variaciones y generalizaciones del problema: búsqueda on-line, búsqueda off-line, búsqueda multipatrón, índices capaces de recuperar palabras o secuencias, búsqueda 2-dimensional (donde incluso el problema debe ser definido formalmente), entre otros.

«Mejores» significa más eficientes en tiempo y/o en espacio y/o en tolerancia al nivel de error. Paralelamente a esto se obtendrá una comprensión más acabada que la actual sobre las estadísticas del problema.