

Seminario de Estructuras de Datos Compactas (10 UD)

Gonzalo Navarro

29 de julio de 2008

Motivación

La brecha entre los tiempos de CPU y los de I/O se ha mantenido creciente durante las últimas décadas. Asimismo han aparecido nuevos niveles en la jerarquía de memoria (caches de tamaño cada vez más considerable). Por ello, se ha hecho cada vez más atractivo el uso de estructuras de datos que ocupen poco espacio, incluso a veces comprimiendo la información sobre la que actúan. Si bien trabajar sobre esta información compacta es más laborioso, el hecho de poder mantenerla en una memoria órdenes de magnitud más rápida la convierte en una alternativa muy conveniente a las implementaciones clásicas. El curso ofrece un puente entre las estructuras de datos y la compresión, pues estudia no sólo la forma de comprimir datos, sino de poder manipularlos en forma comprimida sin necesidad de descomprimirlos. Esta es una rama muy activa y reciente en investigación en algoritmos, y a la vez ofrece herramientas sumamente útiles para desarrollos de alto valor agregado.

Objetivos

El objetivo del curso es entregar a los alumnos una visión de los avances en estructuras de datos compactas y comprimidas, considerando los aspectos teóricos así como su utilidad práctica. El alumno que apruebe el curso conocerá y estará en condiciones de utilizar estructuras de datos de bajo coste de memoria para resolver diversos problemas clásicos, conociendo su complejidad teórica y su aplicabilidad. Asimismo estará capacitado para realizar investigación relacionada con el área.

Metodología

El curso se organiza alrededor de lecturas y exposiciones. Los temas se distribuirán entre los alumnos disponibles. El alumno al que le toque un tema deberá leer los artículos proporcionados por el profesor y los que encuentre por sí mismo, y luego exponerlos en una clase teórica seguida de una sesión de preguntas y discusión grupales (es esperable que el alumno consulte con el profesor varias veces mientras prepara su exposición). Esta exposición debe asimismo considerar aspectos prácticos y resultados experimentales, si los hubiere. El profesor garantizará que la exposición sea correcta y comprensible para los demás alumnos, caso contrario complementará la clase él mismo.

Temario

Introducción. Orientación y metodología del curso. La jerarquía de memoria, velocidades, precio y evolución. Los volúmenes de información a manejar y su evolución. Algunas aplicaciones de ejemplo: Genoma, Yahoo! Compresibilidad y entropía empírica.

Rank y select sobre secuencias de bits, y aplicaciones. Rank y select en tiempo constante usando $n + o(n)$ bits. Rank y select sobre la secuencia comprimida. Aplicaciones a hashing perfecto, sumas parciales y manejo de conjuntos. Secuencias dinámicas (inserción y borrado).

Rank y select sobre secuencias de símbolos, y geometría. El wavelet tree. Soluciones para alfabetos mayores. Consultas sobre rangos bidimensionales. Mínimos en rangos.

Arboles. Representación de paréntesis en preorden. Operaciones de navegación en tiempo constante. Otras representaciones.

Permutaciones y relaciones binarias. Permutación directa e inversa usando $(1 + \epsilon)n \log n$ bits. Alternativa usando wavelet trees. Extensión a relaciones binarias. Representación de Golynski.

Textos y autoíndices El arreglo de sufijos. La transformación Burrows-Wheeler. Búsqueda hacia atrás y el FM-index. La función Ψ y su compresión. El Compressed Suffix Array. Compresión Lempel-Ziv y el LZ-index. Arboles de sufijos comprimidos.

Grafos Fuentes de compresibilidad en grafos de la Web. Vecinos directos y navegación. Autoíndices y vecinos reversos.

Requisitos

Se requerirá CC40A/CC53A.

Evaluación

La evaluación será individual y considerará los siguientes aspectos:

- Proactividad del alumno para investigar sobre su tema.
- Calidad de la exposición, profundidad de la comprensión y respuesta a las preguntas.
- Balance entre teoría y práctica.
- Participación en clase cuando no le toca ser expositor.

Dada la forma de evaluación, se exigirá un 80% de asistencia a clases.

Bibliografía

Artículos científicos de revistas y conferencias internacionales.