

**CC72E SEMINARIO: BUSQUEDA EN LA WEB  
( UD.)**

**Profesor: Ricardo Baeza Y.**

Requisitos :  
Semestre :

### **Resumen**

En este curso estudiamos los problemas principales de buscar en la Web. Comenzamos con las características principales de la Web, su estructura y como se puede modelar. Describimos los principales problemas de los usuarios a través de ejemplos y enumeramos todas las técnicas que existen hoy en día para buscar. En particular explicamos como funcionan las máquinas de búsqueda, su arquitectura, sus limitaciones, como se procesa el texto y como son los índices. A continuación se presentan otras técnicas importantes, como jerarquización de documentos en base a enlaces entre páginas, búsqueda en texto comprimido, búsqueda aproximada y visualización de respuestas. También incluimos el modelo de negocios de los buscadores y las últimas tendencias en XML, metadatos, etc. Estos temas están cubiertos en los capítulos 6, 8 y 13 del libro Modern Information Retrieval (Addison Wesley, 1999).

### **Contenido**

- Introducción
- Desafíos y problemas
- Características de la Web
- Modelando la Web
- Lenguajes para estructurar texto: SGML Y XML
- Máquinas de Búsqueda
  - Arquitecturas centralizadas y distribuidas, interfaces, ranking, crawling, cobertura, índices
- Navegación
  - Directorios, otras herramientas
- Metabuscadore
- Problemas del usuario
- Búsqueda usando hyperlinks
  - Lenguajes de consulta de Web, agentes de software
- Resultados recientes
  - Ranking basado en hyperlinks, búsqueda directa en texto comprimido, lenguajes visuales de consulta y uso de la estructura del texto, visualización de respuestas
- Tendencias y problemas abiertos
  - XML, nuevos modelos de datos, minería de texto

## **Metodología**

Los conceptos teóricos se complementan con ejercicios prácticos para entenderlos y además conocer las diferencias, ventajas y desventajas de distintos buscadores. La evaluación es a través de un trabajo de investigación.

## **Bibliografía**

R. Baeza-Yates & B. Ribeiro-Neto, Modern Information Retrieval, Addison-Wesley 1999.