Tópicos em Recuperação de Informação[1]

Nivio Ziviani

[1] Conjunto de transparências elaborado por Nivio Ziviani e Patrícia Correia

Recuperação de Informação

- Atividade de recuperar itens de informação armazenados em um meio que possa ser recuperado por computador.

- Lida com a representação, armazenamento, organização e acesso a itens de informação.

- Computador produziu uma sociedade que se alimenta da informação.
  - Produção de dados é espantosa: transações financeiras, procedimentos legais, atividades de governo, indústria editorial, etc.

- Como tratar um grande número de documentos (*gigabytes* de dados)?

Desafios

1. Armazenar os dados eficientemente:

   Uso de compressão.

2. Permitir acesso rápido:

   Uso de indexação (eletrônica).

Métodos tradicionais de compressão e busca têm que ser adaptados para responder aos desafios acima. Ex.:

- índice + texto (ambos comprimidos) ocupam menos da metade do arquivo original.
- o tempo para responder a uma consulta é muito menor do que se a base de dados não tivesse sido comprimida!

Bases de documentos

- Armazenamento e recuperação eficientes.
- Compressão $+$ indexação.

$$\Downarrow$$

Soluções isoladas existem há muitos anos, mas o uso combinado das duas técnicas não possível até muito recentemente.

- Imagens fazem parte de documentos.
  - Diagramas e fotografias.
  - Ocupam várias vezes mais espaço do que texto.

Tecnologia

- Muitas técnicas que vamos utilizar no curso foram inventadas e testadas recentemente e somente agora estão sendo colocadas em prática.

- Nenhuma máquina de busca comercial usa todas as técnicas que vamos estudar (ainda).
  - elas seguem o "*hard way*" - *hardware way* - com muito disco e muita memória principal.