Recuperação de Informações por Álgebra Linear Computacional

MAC499 - Projeto de Iniciação Científica

Aluna: Ellen Hidemi Fukuda

Orientador: Paulo José da Silva e Silva

Departamento de Ciência da Computação - IME - USP

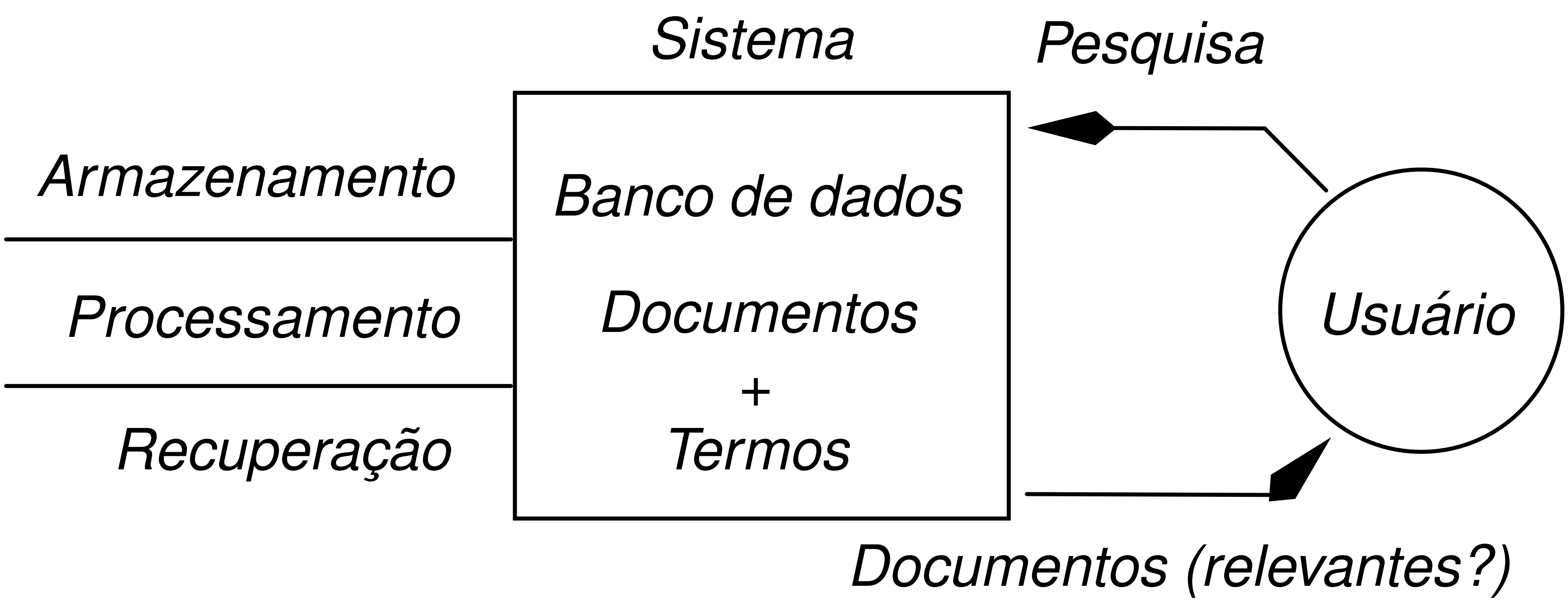
Apoio Financeiro: CNPq

A Iniciação Científica

- Estudo de técnicas de Recuperação de Informações associadas ao modelo vetorial.
- Utilização de ferramentas da Álgebra Linear Computacional, em especial, a Decomposição por Valores Singulares (SVD) e a Fatoração QR.

Recuperação de Informações (IR)

- Métodos eficazes para representação, armazenamento, organização e acesso às informações.



Dificuldades com IR Automático

- Diferentes idiomas.
- Vários tipos de informações: texto, figura, áudio, vídeo.
- Sinônimos (várias palavras com o mesmo significado).
- Polissemia (palavras com diferentes significados).
- Enorme quantidade de documentos.
- Recurso limitado de processamento.

Modelo Vetorial: Termos e Documentos

- Matriz A de termos $\times$ documentos:

$$\begin{array}{cc} & d_j \\ t_i & \begin{bmatrix} a_{11} & \dots & a_{1j} & \dots & a_{1D} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{i1} & \dots & a_{ij} & \dots & a_{iD} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{T1} & \dots & a_{Tj} & \dots & a_{TD} \end{bmatrix} \end{array}$$

- a_{ij} = peso do termo t_i associado ao documento d_j, $1 \leq i \leq T$, $1 \leq j \leq D$.

Modelo Vetorial: Termos e Documentos (Cont.)

- Definições possíveis para a_{ij}: variável booleana, freqüência do termo no documento, funções envolvendo logaritmos, etc. As colunas da matriz A podem ou não ser normalizadas.
- Se o termo t_i não estiver relacionado ao documento d_j, então $a_{ij} = 0$.
- Em geral, o número de termos relacionados a um documento é razoavelmente pequeno. A matriz é, portanto, esparsa.

Modelo Vetorial: Pesquisas

- Cada pesquisa é definida como um vetor $q = (q_1, ..., q_T)^T$.
- Medida de similaridade entre uma pesquisa q e um documento $d_j = (a_{1j}, ..., a_{Tj})^T$:

$$\cos(\theta_j) = \frac{d_j^T q}{\|d_j\|_2 \|q\|_2} = \frac{\sum_{i=1}^{T} a_{ij} q_i}{\sqrt{\sum_{i=1}^{T} a_{ij}^2} \sqrt{\sum_{i=1}^{T} q_i^2}}$$

- Seja L um limiar definido. Se $\cos(\theta_j) > L$, então d_j é um documento relevante para a pesquisa q.

Redução do Posto da Matriz (LSI)

- LSI (*Latent Semantic Indexing*): É baseado no modelo vetorial e utiliza-se da matriz de termos $\times$ documentos com posto reduzido.
- A redução do posto permite remover algumas informações não-pertinentes.

Decomposição SVD

- Decomposição SVD de $A \in \mathbb{R}^{T \times D}$:

$$A = U \Sigma V^T,$$

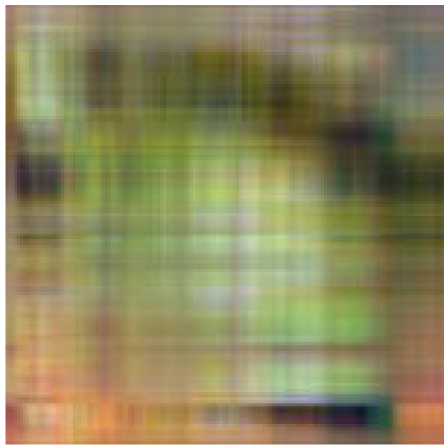
onde $U \in \mathbb{R}^{T \times T}$ e $V \in \mathbb{R}^{D \times D}$ são matrizes ortogonais e $\Sigma \in \mathbb{R}^{T \times D}$ é uma matriz diagonal cujos elementos são os valores singulares $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\min(T,D)}$.

- O posto r_A da matriz A é igual ao número de valores singulares não nulos.

Propriedade do SVD

- Se $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$, então A_k é a melhor aproximação de A de posto k.
- Cada eixo da hiperelipse associada à matriz A fornece uma informação proporcional a σ_i.
- Escolher um k apropriado não é simples. Usualmente é definido através de experimentos.

Analogia com Compressão de Imagens



k = 3

k = 7

k = 15

k = 40

k = 75

r_A = 150

Outros Estudos

- Uso da fatoração QR no contexto de IR.
- Criação de *thesaurus*: *clustering*, comparação entre termos e entre documentos.
- Operações com vetores de pesquisas: expansão da pesquisa.
- Gerenciamento de coleções dinâmicas.

Mais Informações

- Página de MAC499: http://www.linux.ime.usp.br/~hidemi/mac499.
- E-mail: ellen at ime.usp.br