Recuperação de Informações por Álgebra Linear Computacional

Aluna: Ellen Hidemi Fukuda / Orientador: Paulo José da Silva e Silva

Ciência da Computação - IME

E-mails: ellen at ime.usp.br

e rsilva at ime.usp.br

1 Recuperação de Informações (IR)

Com a evolução de bibliotecas digitais e o crescimento exponencial da quantidade de documentos disponíveis na Internet, tornaram-se necessários métodos eficazes para o armazenamento, o processamento e a recuperação de informações. Tais métodos podem ser aplicados a um grande banco de dados, quando implementados em sistemas de alta performance. Um exemplo de sistema de grande escala conhecido atualmente é o Google. Seus usuários definem perguntas e o sistema fornece conjuntos de documentos relacionados a elas através de um processamento de dados.

Técnicas automáticas desse tipo, no entanto, não são fáceis de implementar. Problemas associados à ambigüidade da linguagem natural, aos diversos idiomas existentes e aos tipos de informações (texto, figuras, áudio e vídeo) motivam pesquisadores a estudarem diversas maneiras de contorná-los. Indexar grandes quantidades de dados usando recursos limitados de processamento e retornar documentos realmente relevantes às pesquisas são ainda grandes desafios.

Uma tecnologia de recuperação de informações desenvolvida recentemente é o LSI (*Latent Semantic Indexing*), baseada no modelo clássico vetorial. Nesse modelo, as informações são armazenadas em uma matriz, onde cada coluna representa um documento e cada linha está associada a um termo do "dicionário". A pesquisa do usuário ao banco de dados é representada por um vetor. A identificação de documentos relevantes à pesquisa e a atualização do banco de dados é feita utilizando-se algoritmos conhecidos da Álgebra Linear Computacional, em especial, a decomposição por valores singulares (SVD).

2 Sistema de IR

- Um sistema de IR consiste em um programa que facilita usuários a encontrarem informações desejadas. Seu principal objetivo é de minimizar o trabalho de um usuário durante a pesquisa. Esse trabalho pode ser expresso pelo tempo gasto durante a busca, além da qualidade e quantidade das informações obtidas.
- Para uma certa pesquisa, temos dois tipos de documentos no banco de dados - os relevantes e os não relevantes - dentre os quais alguns serão retornados. A figura 1 ilustra o conjunto total de documentos e suas divisões para uma pesquisa.

Figura 1: Possíveis resultados de busca em um sistema de IR.

- Um sistema de IR é considerado eficiente se tiver um alto *retorno*, ou seja, se o número de documentos relevantes retornados sobre os relevantes possíveis for grande, e se a *precisão* for alta, isto é, se a maioria dos documentos retornados for relevante. O *retorno* e a *precisão*, no entanto, se conflitam: a diminuição de uma geralmente implica no aumento de outra. O desafio de um sistema IR está justamente em tentar conciliá-las.

3 O Modelo Vetorial

- O banco de dados é representado por uma matriz A de termos $\times$ documentos:

$$A = \begin{bmatrix} a_{11} & \dots & a_{1j} & \dots & a_{1D} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{i1} & \dots & a_{ij} & \dots & a_{iD} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{T1} & \dots & a_{Tj} & \dots & a_{TD} \end{bmatrix}, \qquad 1 \leq i \leq T,\ 1 \leq j \leq D.$$

- T é o número de termos e D é o número de documentos do banco de dados do sistema.
- Um termo t_i está associado à i-ésima linha de A e um documento d_j é representado pela j-ésima coluna de A, ou seja, $d_j = (a_{1j}, ..., a_{Tj})^T$.
- Cada elemento a_{ij} da matriz indica o quanto o termo t_i está relacionado ao documento d_j. Tais elementos podem ser definidos de diversas maneiras: variável booleana, freqüência do termo no documento, funções envolvendo logarítmos, etc. Além disso, as colunas de A podem ou não ser normalizadas. Note que se t_i não tiver relação com d_j, então $a_{ij} = 0$.
- Em um sistema real de IR, cada documento está associado a uma quantia relativamente pequena de termos. A matriz A é, portanto, esparsa. Seu armazenamento deve-se dar de forma eficiente e técnicas específicas devem ser utilizadas.
- Cada pesquisa do usuário é definida como um vetor $p = (p_1, ..., p_T)^T$, onde cada elemento p_i é o peso que um termo t_i possui na pesquisa p. Da mesma forma que os elementos de A, p_i pode ser definida de inúmeras maneiras e p pode ou não ter norma igual a um.
- A similaridade de uma pesquisa p e um documento $d_j = (a_{1j}, ..., a_{Tj})^T$ é dada pelo cosseno do ângulo formado por esses vetores, ou seja:

$$\text{sim}(p, d_j) \doteq \cos(\theta_j) = \frac{d_j^T p}{\|d_j\|_2 \|p\|_2} = \frac{\sum_{i=1}^{T} a_{ij} p_i}{\sqrt{\sum_{i=1}^{T} a_{ij}^2} \sqrt{\sum_{i=1}^{T} p_i^2}}$$

- Utilizando-se da medida acima, um documento d_j será considerado relevante para uma pesquisa p se $\text{sim}(p, d_j) > L$, para um certo limiar L definido pelo sistema. Definir esse L requer inúmeros experimentos e depende do tamanho e tipo de informações contidas no banco de dados.

4 Decomposição SVD

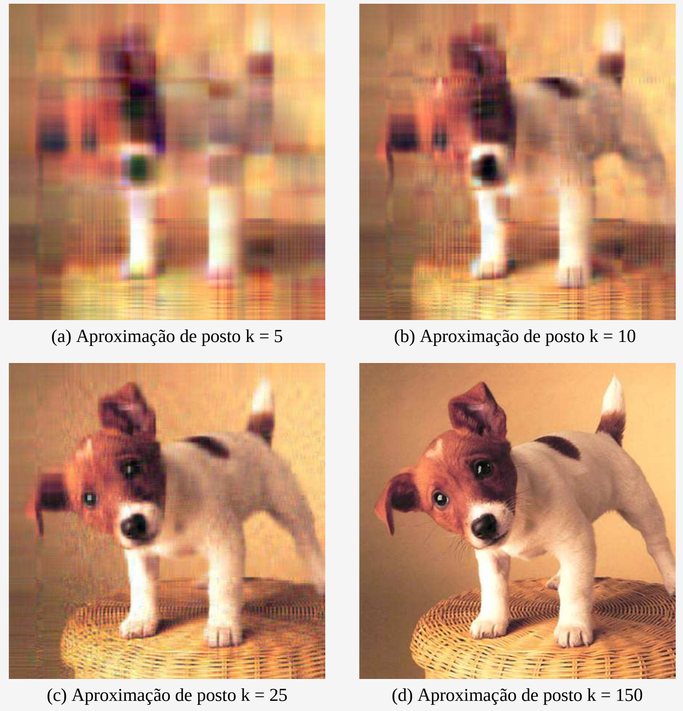
- A decomposição SVD de $A \in \mathbb{R}^{T \times D}$ é dada por $A = U \Sigma V^T$, onde $U \in \mathbb{R}^{T \times T}$ e $V \in \mathbb{R}^{D \times D}$ são matrizes ortogonais e $\Sigma \in \mathbb{R}^{T \times D}$ é uma matriz diagonal cujos elementos são os valores singulares $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_{\min(T,D)}$.
- No universo dos números reais, a existência do SVD está associada ao fato de que a imagem de uma esfera unitária sob uma matriz $A \in \mathbb{R}^{m \times n}$ é uma *hiperelipse* no $\mathbb{R}^m$. A esfera unitária é definida por vetores ortonormais $v_i \in \mathbb{R}^n$. Aplicando-se a matriz A, teremos uma hiperelipse com eixos $u_i \in \mathbb{R}^m$ com comprimento σ_i. Desse modo, $Av_i = \sigma_i u_i$, $i = 1, ..., n$.
- O número de valores singulares não nulos de A é igual ao posto r_A de A.
- A decomposição SVD de A também pode ser escrita da seguinte forma:

$$A = U \Sigma V^T = \sum_{i=1}^{r_A} \sigma_i u_i v_i^T,$$

com u_i e v_i sendo as i-ésimas colunas de U e de V, respectivamente.

5 Latent Semantic Indexing

- Um banco de dados de um sistema de recuperação de informações representado pela matriz de termos $\times$ documentos pode ser construído por um longo tempo, por inúmeras pessoas de opiniões e conhecimentos distintos. Todas essas incertezas são refletidas na matriz e um bom meio de contornar esse problema é usando a decomposição SVD da matriz.
- A idéia do *Latent Semantic Indexing* (LSI) é substituir a matriz original A por outra que possuísse menos incertezas (ou "erros"). Tal matriz, A_k, será uma aproximação de A, de posto $k < r_A$, definida por $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$. Observe que se $k = r_A$, obtemos a matriz original A.
- Escolhendo-se adequadamente o valor de k, pode-se obter melhores resultados durante a pesquisa. Além disso, com a diminuição do posto, o tempo de processamento é menor, bem como o espaço necessário para o armazenamento das informações.
- A idéia do LSI é análoga a da compressão de imagens usando SVD. Uma imagem pode ser representada por uma matriz de posto p. A hiperelipse associada a ela possui eixos de tamanho σ_i, para algum i. Cada eixo, por sua vez, fornece uma informação proporcional a σ_i. Assim, se armazenarmos essa imagem em uma matriz de posto $k < p$ (i.e., se usarmos a aproximação da matriz de posto reduzido), a imagem originada desta será menos nítida que a original, já que foram retirados $p - k$ eixos da hiperelipse. A figura 2 ilustra essa idéia. O posto da matriz original é $p = 430$. Note que um posto $k = 150$ já nos oferece uma boa imagem.



(a) Aproximação de posto k = 5

(b) Aproximação de posto k = 10

(c) Aproximação de posto k = 25

(d) Aproximação de posto k = 150

Figura 2: Compressão de imagens usando SVD.

MAC499 - Trabalho de Formatura Supervisionado: Iniciação científica parcialmente financiada pelo CNPq. Julho de 2003 a Dezembro de 2004.