

MAC 499 - Trabalho de Formatura Supervisionado

Recuperação de Informações em Bancos de Dados Textuais

Aluna: Marcela Ortega Garcia

Orientador: Prof. Dr. João Eduardo Ferreira

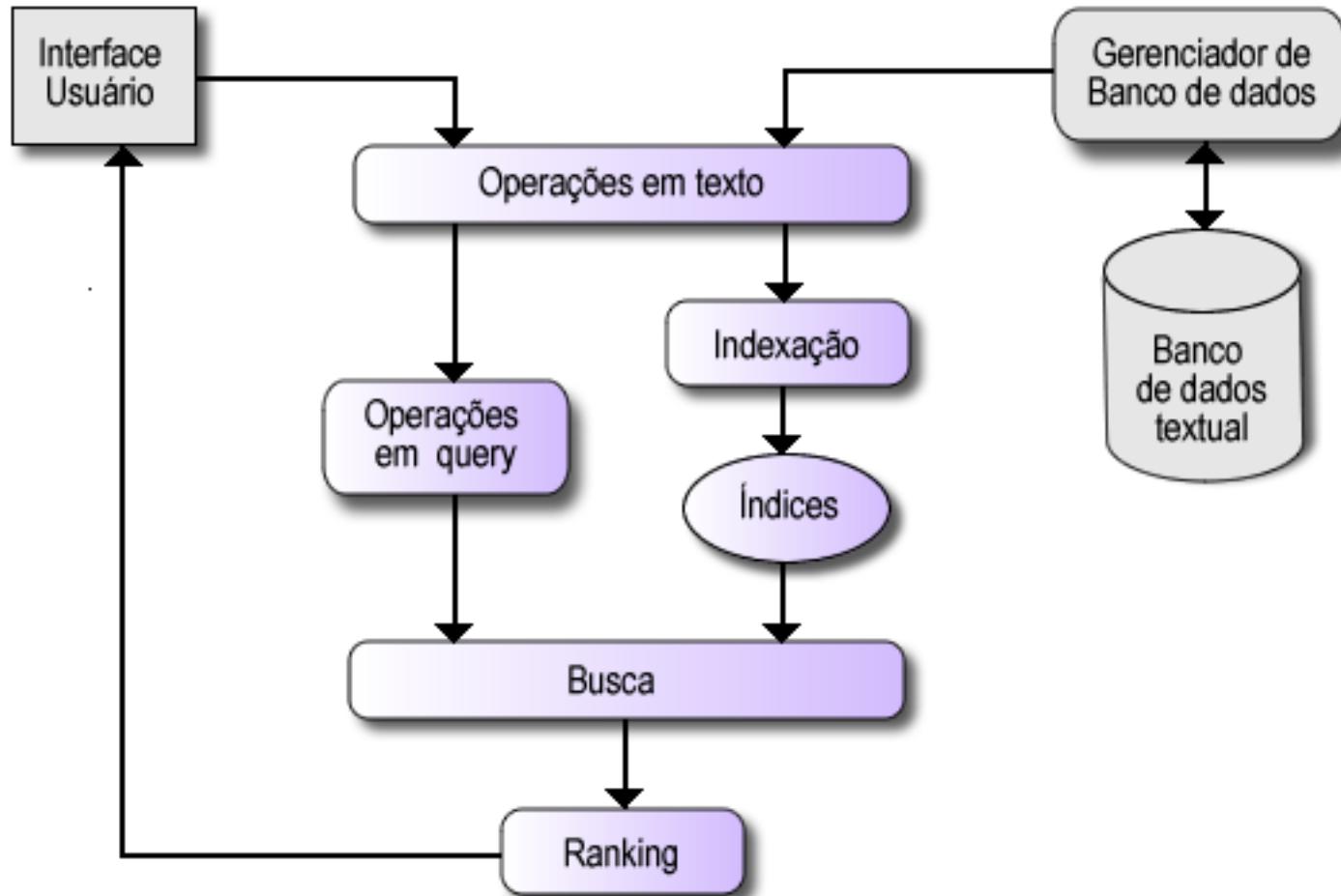
Introdução

- Armazenamento e recuperação eficientes
- Dados estruturados
 - Tabelas em bancos de dados relacionais
- Dados não-estruturados
 - Textos
- Aumento do volume de dados → Novas técnicas

Recuperação de Informações

- Representação, armazenamento, organização e acesso a informações
- Dados x Informações
 - Linguagem ambígua
 - Relevância: centro da recuperação de informações
- Índices - Palavras que representam o texto
- Consulta - Tradução em palavras-chave
 - “Quais as queixas dos pacientes com DMC?” → queixas DMC

O processo de RI



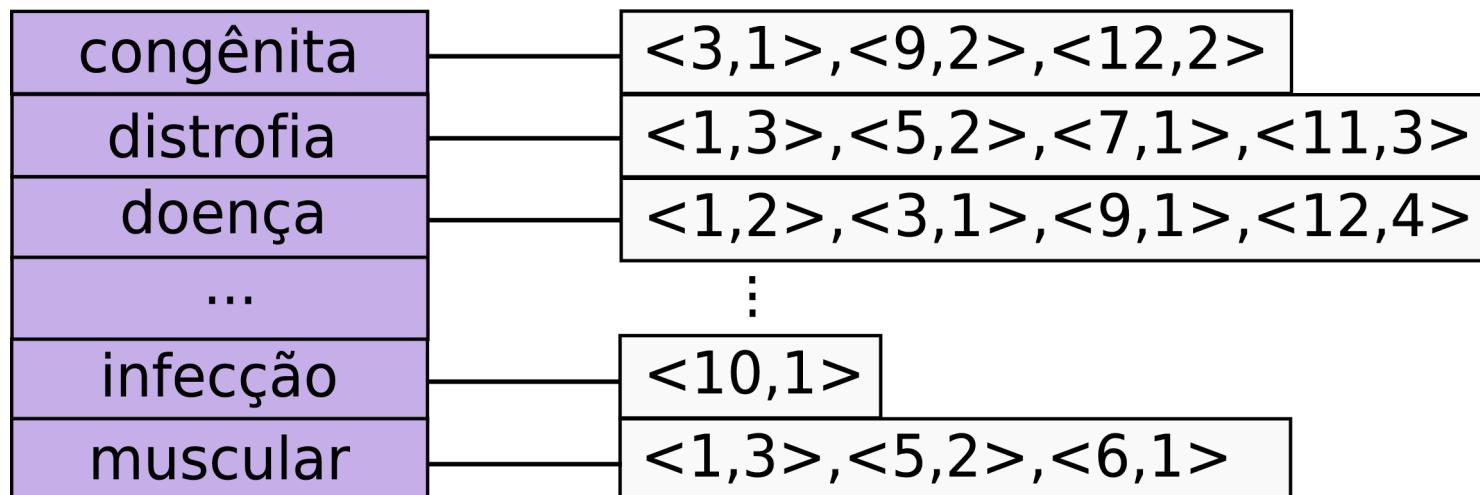
Operações em texto

- Análise léxica - Identificação de palavras
 - arco-íris → arco e íris ou arco-íris?
 - usuario@email.com → usuario, email, com ?
- Eliminação de *stopwords*
 - a, o, um, de, para
- *Stemming* - Raiz gramatical
 - “recuperação”, “recuperam”, “recuperado” → recuperar

Indexação

- Arquivos invertidos

- Orientado à palavra
- Vocabulário e listas de ocorrências
- $\langle d, f_{d,t} \rangle$: <documento d , frequência do termo t no documento d >



Modelos de RI

- Definido pelo algoritmo de ranqueamento
- Modelo booleano
 - Pesos booleanos às palavras: presença ou ausência em um documento.
 - Consulta com formato de expressão booleana
 - *Exemplo:* doença AND muscular NOT herança

$$sim(d_j, q) = \begin{cases} 1 & \text{Se a expressão é válida para } d_j \\ 0 & \text{Caso contrário.} \end{cases}$$

Modelo vetorial

- Pesos não-binários → Grau de relevância
- Pesos relacionados aos documentos e às consultas

$$w_{i,j} = f_{i,j} \times \log \frac{N}{n_i}$$

$$w_{i,q} = (0.5 + 0.5f_{i,q}) \times \log \frac{N}{n_i}$$

$f_{i,j} \rightarrow$ frequência normalizada do termo k_i no documento d_j

$N \rightarrow$ número de documentos na coleção

$n_i \rightarrow$ número de documentos em que o termo k_i aparece

$f_{i,q} \rightarrow$ frequência normalizada do termo k_i na consulta q

Modelo vetorial

- Vetores de pesos

$$\vec{d}_j = (w_{1,j}, w_{2,j}, \dots, w_{t,j})$$

$$\vec{q} = (w_{1,q}, w_{2,q}, \dots, w_{t,q})$$

- Similaridade: correlação entre vetores

$$sim(\vec{d}_j, \vec{q}) = \frac{\vec{d}_j \cdot \vec{q}}{|\vec{d}_j| \times |\vec{q}|}$$

Ferret

- Biblioteca de indexação e busca
- Baseada no *Lucene*, escrita em *Ruby*
- Combinação do modelo booleano com o vetorial
- Ferret :: Analysis :: Analyzer
- Ferret :: QueryParser
 - AND, OR, NOT
- Ferret :: Search :: Searcher

CEGH

- Sistema *Web* do Centro de Estudos do Genoma Humano
- Campos do tipo texto
 - Cadastro de paciente
 - Anotação de consulta
- Necessidade de busca: índice único

CEGH

● Cadastro de paciente

Cadastro de Paciente

Grupo * Data cadastro

Externo

Caso índice?

Nome *
João da Silva

Sexo *
Masculino

Data nascimento Idade

[Endereço](#) [Contato](#) [Ocultar observações](#)

Observações

* Campos obrigatórios

CEGH

● Anotação de consulta

Criando anotação de consulta

Paciente

Registro: P001176 - Nome: João da Silva

[Informações Adicionais](#)

Observações

Solicitar exames

Deixe o prazo previsto em branco para assumir o tempo padrão do exame

Ordem de execução	Exame	Prazo previsto	Adicionar
	<input type="text"/>	<input type="text"/>	

Próximos passos

O prazo previsto é opcional

Ordem de execução	Ação	Complementação	Prazo previsto	Adicionar
	<input type="text"/>	<input type="text"/>	15/11/2009	

[Salvar](#)

[Consultas anteriores](#)

CEGH

● Tela de Busca

Busca

fendas AND parciais

Buscar

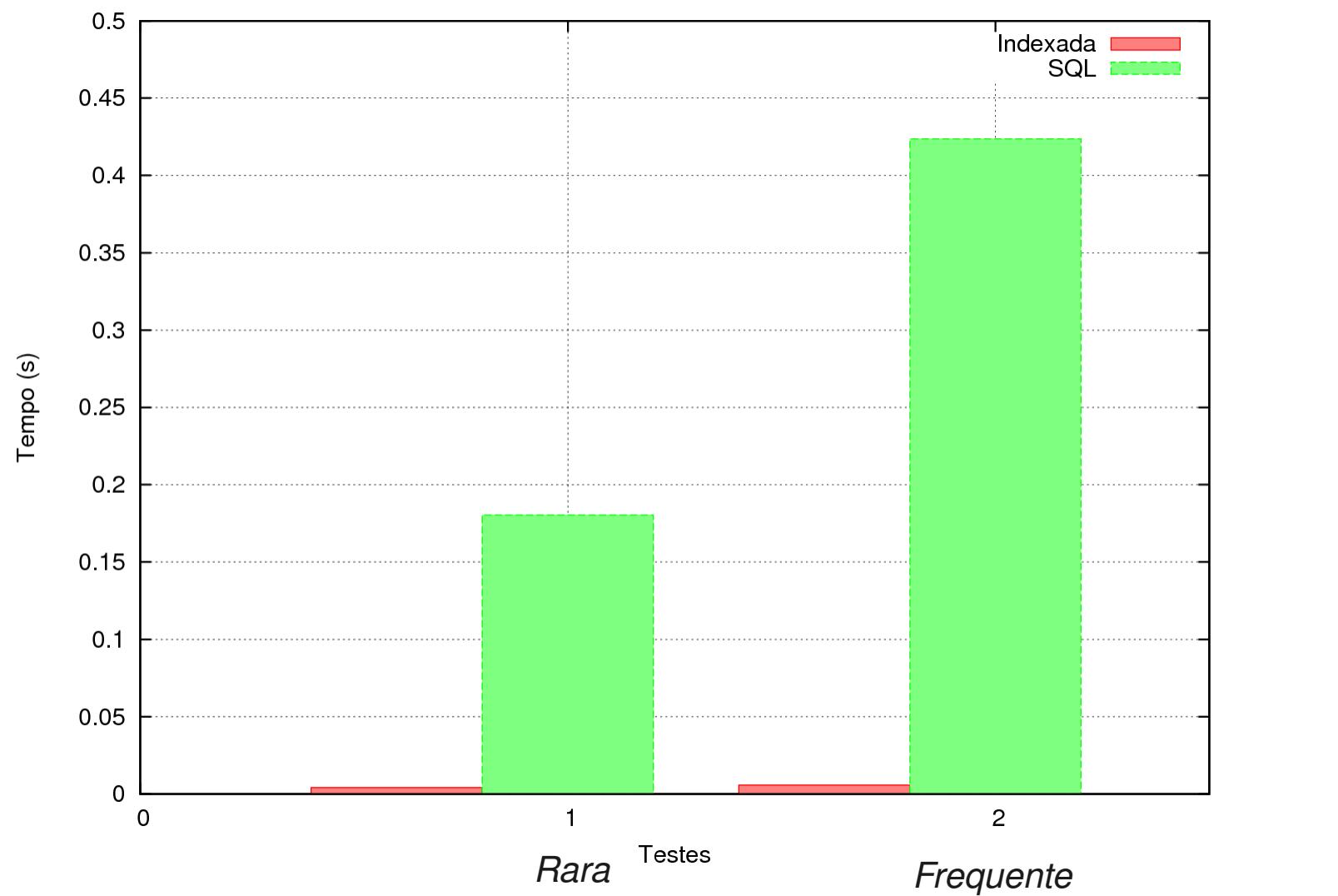
● Tela de Resultado

Listando resultado de busca

Resultados para: **fendas AND parciais**

Tipo	Paciente	
Anotação de consulta	Rafael de Souza	<button>Ver</button>
Cadastro de paciente	Rafael de Souza	<button>Ver</button>
Cadastro de paciente	Joana dos Santos	<button>Ver</button>
Anotação de consulta	João da Silva	<button>Ver</button>
Anotação de consulta	Maria Aparecida	<button>Ver</button>
Cadastro de paciente	José Oliveira	<button>Ver</button>
Cadastro de paciente	João da Silva	<button>Ver</button>
Anotação de consulta	Roberto Pereira	<button>Ver</button>

Resultados



Dúvidas?

Referências:

- R. Baeza-Yates, B. Ribeiro-Neto - *Modern information retrieval*
- J. Zobel, A. Moffat. - *Inverted files for text search engines*
- E. Bertino, K.L. Tan, B.C. Ooi, R. Sacks-Davis, J. Zobel, e B. Shidlovsky - *Indexing techniques for advanced database systems.*
- Apache Lucene - <http://lucene.apache.org>
- Ferret - <http://www.davebalmain.com>

Contato: marcela.ortega@usp.br