Organização e Recuperação de Informação – GSI521

Prof. Dr. Rodrigo Sanches Miani – FACOM/UFU

Índice invertido

Organização e Recuperação de Informação (GSI521)

Laboratório 1 - Pergunta

- Como os termos foram associados aos documentos e as ocorrências?
- Como as buscas foram feitas?

Introdução

- Um **índice invertido** é um mecanismo orientado a palavras para a indexação de uma coleção de texto a fim de acelerar a tarefa de busca;
- A estrutura do índice invertido é composta por dois elementos: o **vocabulário** (dicionário) e as **ocorrências**;
- A maioria dos sistemas modernos de RI funciona usando um índice invertido.

Introdução

- **Vocabulário**: conjunto de todas as palavras diferentes no texto;
- Para cada palavra do vocabulário, o índice armazena os documentos que contêm esta palavra;
- Por isso o nome índice invertido, pois podemos reconstruir o texto através do índice;
- Esse é o principal e o mais antigo índice para recuperação de informação.

Exemplo

Documento 1: eu encenei julio cesar eu fui morto no capitolio brutus me matou

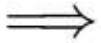
Documento 2: que assim seja com cesar o nobre brutus vos disse que cesar era ambicioso

Exemplo – Passo 1) Identificar os termos e relacionar com os documentos

termo	docID
eu	1
encenei	1
júlio	1
césar	1
eu	1
fui	1
morto	1
no	1
capitólio	1
brutus	1
me	1
matou	1
que	2
assim	2
seja	2
com	2
césar	2
o	2
nobre	2
brutus	2
vos	2
disse	2
que	2
césar	2
era	2
ambicioso	2

Exemplo – Passo 2) Ordenar os termos

termo	docID
eu	1
encenei	1
júlio	1
césar	1
eu	1
fui	1
morto	1
no	1
capitólio	1
brutus	1
me	1
matou	1
que	2
assim	2
seja	2
com	2
césar	2
o	2
nobre	2
brutus	2
vos	2
disse	2
que	2
césar	2
era	2
ambicioso	2

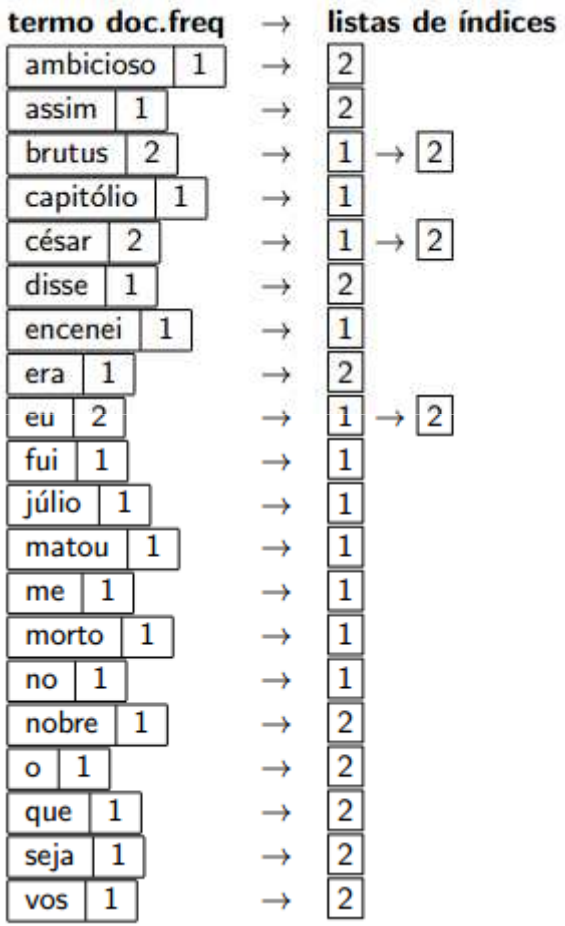


termo	docID
ambicioso	2
assim	2
brutus	1
brutus	2
capitólio	1
com	2
césar	1
césar	2
césar	2
disse	2
encenei	1
era	2
eu	1
eu	1
fui	1
júlio	1
matou	1
me	1
morto	1
no	1
nobre	2
o	2
que	2
que	2
seja	2
vos	2

Exemplo – Passo 3) Determinar as frequências de documentos

termo	doc.freq
ambicioso	1
assim	1
brutus	2
capitólio	1
césar	2
disse	1
encenei	1
era	1
eu	2
fui	1
júlio	1
matou	1
me	1
morto	1
no	1
nobre	1
o	1
que	1
seja	1
vos	1

Exemplo – Passo 4) Criar a lista com os documentos



Exemplo de buscas usando o índice invertido

- Considere a busca BRUTUS and MATOU
  - Localizar BRUTUS no dicionário (pode-se utilizar um termID para cada palavra no vocabulário);
  - Recuperar a lista de índices de BRUTUS
  - Localizar MATOU;
  - Recuperar a lista de índices de MATOU;
  - Fazer a intersecção entre as duas listas.

Estrutura de dados - Dicionario

- Cada termo será cadastrado na forma de um Termo no Dicionario e receberá um termID (um número inteiro);
- Cada documento também receberá um docID;
- Criação de listas encadeadas associando cada termID aos respectivos docIDs;
- Dado um termID, o Dicionario será consultado para obter as listas encadeadas associadas a cada termID.

Pesquisa por termo

- A pesquisa no vocabulário pode ser efetuada utilizando qualquer estrutura de dados apropriada, como:
  - Hashing
  - Árvores
- Alguns sistemas de ORI usam Hashing outros usam Árvores;
- Ambos possuem vantagens e desvantagens;
  - Hashing: tempo de busca / reconstrução das tabelas ao longo do tempo;
  - Árvores: buscas com prefixos / rebalanceamento

Índices invertidos completos

- O índice discutido anteriormente não é adequado para responder a consultas com:
  - Frases
  - Por proximidade
- Não contém informações sobre onde exatamente no documento cada palavra aparece.

Índices invertidos completos

- Para isso, precisamos adicionar as posições de cada palavra em cada documento ao índice;
- Essas posições podem se referir a palavras ou caracteres.

Índices invertidos completos

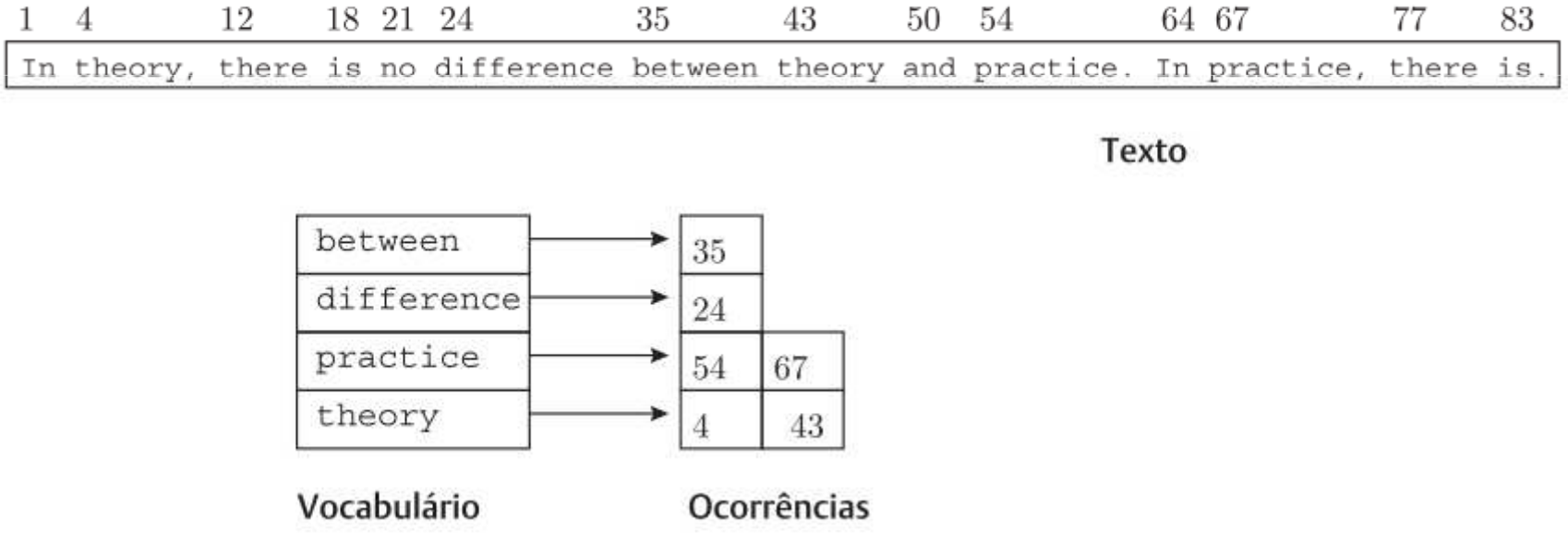
- **Posições de palavras** - a posição *i* refere-se à *i-ésima* palavra;
- Simplificam a busca por frases e consultas por proximidade.

Índices invertidos completos

- **Posições de carateres**- a posição *i* refere-se ao *i-ésimo* caractere;
- Facilitam o acesso direto às posições de texto correspondentes, por exemplo, para exibir snippets de texto.

Índices invertidos completos - Exemplo

Considere o caso para um único texto, com cada ocorrência de termo assinalada pela sua posição de caractere:



Índices invertidos completos - Exemplo

No caso de vários documentos, precisamos armazenar uma lista de ocorrências para cada par termo-documento. Considere o seguinte vocabulário:

D0 = "it is what it is"

D1 = "what is it"

D2 = "it is a banana"

Índices invertidos completos - Exemplo

"a": {(2, 2)}
"banana": {(2, 3)}
"is": {(0, 1), (0, 4), (1, 1), (2, 1)}
"it": {(0, 0), (0, 3), (1, 2), (2, 0)}
"what": {(0, 2), (1, 0)}

Índices invertidos completos - Exercício

- Como ficaria o índice invertido completo (palavras) para a seguinte coleção de documentos:

  Documento 1: eu encenei julio cesar eu fui morto no capitolio brutus me matou

  Documento 2: que assim seja com cesar o nobre brutus vos disse que cesar era ambicioso

Índices invertidos completos – Frequência de termos

- Também podemos construir um índice invertido completo usando a frequência de termos ao invés da posição de palavras ou caracteres:

D0 = "it is what it is"

D1 = "what is it"

D2 = "it is a banana"

Índices invertidos completos – Frequência de termos

"a": {(2,1)}

"banana": {(2,1)}

"is": {(0,2), (1,1), (2,1)}

"it": {(0,2), (1,1), (2,1)}

"what": {(0,1), (1,1)}