# Arquivos invertidos

# Arquivos invertidos

- É um "mecanismo" que utiliza palavras para indexar uma coleção de documentos
  - a fim de facilitar a busca e a recuperação
- Estruturas de um arquivo invertido
  - Vocabulário
    - É o conjunto de todas as palavras distintas no texto
  - Ocorrências
    - Lista que contém toda a informação necessária sobre cada palavra do vocabulário
    - E.g., documentos onde a palavra aparece, sua posição no texto, freqüência, etc…

# Arquivos Invertidos
## Exemplo

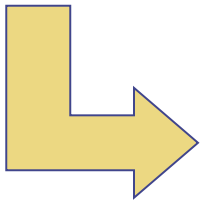
Base de Documentos

| Documento | Texto |
|---|---|
| 1 | Pease porridge hot, pease porridge cold |
| 2 | Pease porridge in the pot |
| 3 | Nine days cold |
| 4 | Some like it hot, some like it cold |
| 5 | Some like it in the pot |
| 6 | Nine days old |

# Arquivos Invertidos
# Exemplo

**Base de Documentos**

| Documento | Texto |
|---|---|
| 1 | Pease porridge hot, pease porridge cold |
| 2 | Pease porridge in the pot |
| 3 | Nine days cold |
| 4 | Some like it hot, some like it cold |
| 5 | Some like it in the pot |
| 6 | Nine days old |



**Arquivo Invertido**

| No | Termo | Docs |
|---|---|---|
| 1 | cold | 1, 4 |
| 2 | days | 3, 6 |
| 3 | hot | 1, 4 |
| 4 | in | 2, 5 |
| 5 | it | 4, 5 |
| 6 | like | 4, 5 |
| 7 | nine | 3, 6 |
| 8 | old | 3, 6 |
| 9 | pease | 1, 2 |
| 10 | porridge | 1, 2 |
| 11 | pot | 2, 5 |
| 12 | some | 4, 5 |
| 13 | the | 2, 5 |

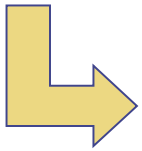
Vocabulário

Listas de documentos
onde termo aparece

# Arquivos Invertidos
## Exemplo

**Base de Documentos**

| Documento | Texto |
|---|---|
| 1 | Pease porridge hot, pease porridge cold |
| 2 | Pease porridge in the pot |
| 3 | Nine days cold |
| 4 | Some like it hot, some like it cold |
| 5 | Some like it in the pot |
| 6 | Nine days old |



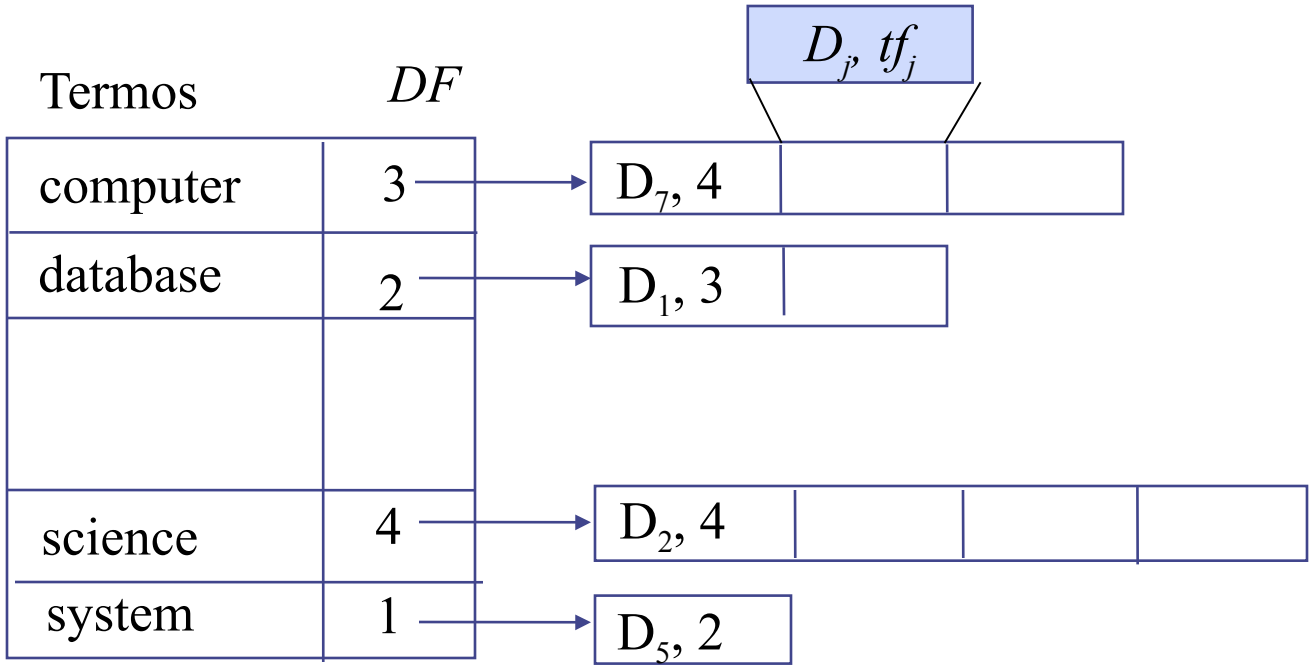
**Arquivo Invertido**

| No | Termo | (Docs; Pos) |
|---|---|---|
| 1 | cold | (1;6), (4;8) |
| 2 | days | (3;2), (6;2) |
| 3 | hot | (1;3), (4;4) |
| 4 | in | (2;3), (5;4) |
| 5 | it | (4;3,7), (5;3) |
| 6 | like | (4;2,6), (5;2) |
| 7 | nine | (3;1), (6;1) |
| 8 | old | (3;3), (6;3) |
| 9 | pease | (1;1,4), (2;1) |
| 10 | porridge | (1;2,5), (2;2) |
| 11 | pot | (2;5), (5;6) |
| 12 | some | (4;1,5), (5;1) |
| 13 | the | (2;4), (5;5) |

Vocabulário

Ocorrências e posições

# Arquivo Invertido com TF-IDF




- Entrada do vocabulário armazena a freqüência do termo na base - DF

- Cada ocorrência indica o documento onde o termo aparece e a freqüência do termo no documento - TF

# Arquivo Invertido com TF-IDF
## Construção

**1.** Texto dos documentos é pré-processado para extrair os termos relevantes, que são armazenados de forma seqüencial juntamente com o identificador dos documentos (Doc#)

Doc 1

Now is the time for all good men to come to the aid of their country

Doc 2

It was a dark and stormy night in the country manor. The time was past midnight

| Term | Doc # |
|---|---|
| now | 1 |
| is | 1 |
| the | 1 |
| time | 1 |
| for | 1 |
| all | 1 |
| good | 1 |
| men | 1 |
| to | 1 |
| come | 1 |
| to | 1 |
| the | 1 |
| aid | 1 |
| of | 1 |
| their | 1 |
| country | 1 |
| it | 2 |
| was | 2 |
| a | 2 |
| dark | 2 |
| and | 2 |
| stormy | 2 |
| night | 2 |
| in | 2 |
| the | 2 |
| country | 2 |
| manor | 2 |
| the | 2 |
| time | 2 |
| was | 2 |
| past | 2 |
| midnight | 2 |

# Arquivo Invertido com TF-IDF
## Construção

**2.** O arquivo gerado é ordenado lexicograficamente (=ordem alfabética)

| Term | Doc # |
|---|---|
| now | 1 |
| is | 1 |
| the | 1 |
| time | 1 |
| for | 1 |
| all | 1 |
| good | 1 |
| men | 1 |
| to | 1 |
| come | 1 |
| to | 1 |
| the | 1 |
| aid | 1 |
| of | 1 |
| their | 1 |
| country | 1 |
| it | 2 |
| was | 2 |
| a | 2 |
| dark | 2 |
| and | 2 |
| stormy | 2 |
| night | 2 |
| in | 2 |
| the | 2 |
| country | 2 |
| manor | 2 |
| the | 2 |
| time | 2 |
| was | 2 |
| past | 2 |
| midnight | 2 |

| Term | Doc # |
|---|---|
| a | 2 |
| aid | 1 |
| all | 1 |
| and | 2 |
| come | 1 |
| country | 1 |
| country | 2 |
| dark | 2 |
| for | 1 |
| good | 1 |
| in | 2 |
| is | 1 |
| it | 2 |
| manor | 2 |
| men | 1 |
| midnight | 2 |
| night | 2 |
| now | 1 |
| of | 1 |
| past | 2 |
| stormy | 2 |
| the | 1 |
| the | 1 |
| the | 2 |
| the | 2 |
| their | 1 |
| time | 1 |
| time | 2 |
| to | 1 |
| to | 1 |
| was | 2 |
| was | 2 |

# Arquivo Invertido com TF-IDF
## Construção

**3.** Múltiplas entradas do termo para o mesmo documento são então agrupadas, e a informação da freqüência é adicionada

| Term | Doc # |
|---|---|
| a | 2 |
| aid | 1 |
| all | 1 |
| and | 2 |
| come | 1 |
| country | 1 |
| country | 2 |
| dark | 2 |
| for | 1 |
| good | 1 |
| in | 2 |
| is | 1 |
| it | 2 |
| manor | 2 |
| men | 1 |
| midnight | 2 |
| night | 2 |
| now | 1 |
| of | 1 |
| past | 2 |
| stormy | 2 |
| the | 1 |
| the | 1 |
| the | 2 |
| the | 2 |
| their | 1 |
| time | 1 |
| time | 2 |
| to | 1 |
| to | 1 |
| was | 2 |
| was | 2 |

| Term | Doc # | Freq |
|---|---|---|
| a | 2 | 1 |
| aid | 1 | 1 |
| all | 1 | 1 |
| and | 2 | 1 |
| come | 1 | 1 |
| country | 1 | 1 |
| country | 2 | 1 |
| dark | 2 | 1 |
| for | 1 | 1 |
| good | 1 | 1 |
| in | 2 | 1 |
| is | 1 | 1 |
| it | 2 | 1 |
| manor | 2 | 1 |
| men | 1 | 1 |
| midnight | 2 | 1 |
| night | 2 | 1 |
| now | 1 | 1 |
| of | 1 | 1 |
| past | 2 | 1 |
| stormy | 2 | 1 |
| the | 1 | 2 |
| the | 2 | 2 |
| their | 1 | 1 |
| time | 1 | 1 |
| time | 2 | 1 |
| to | 1 | 2 |
| was | 2 | 2 |

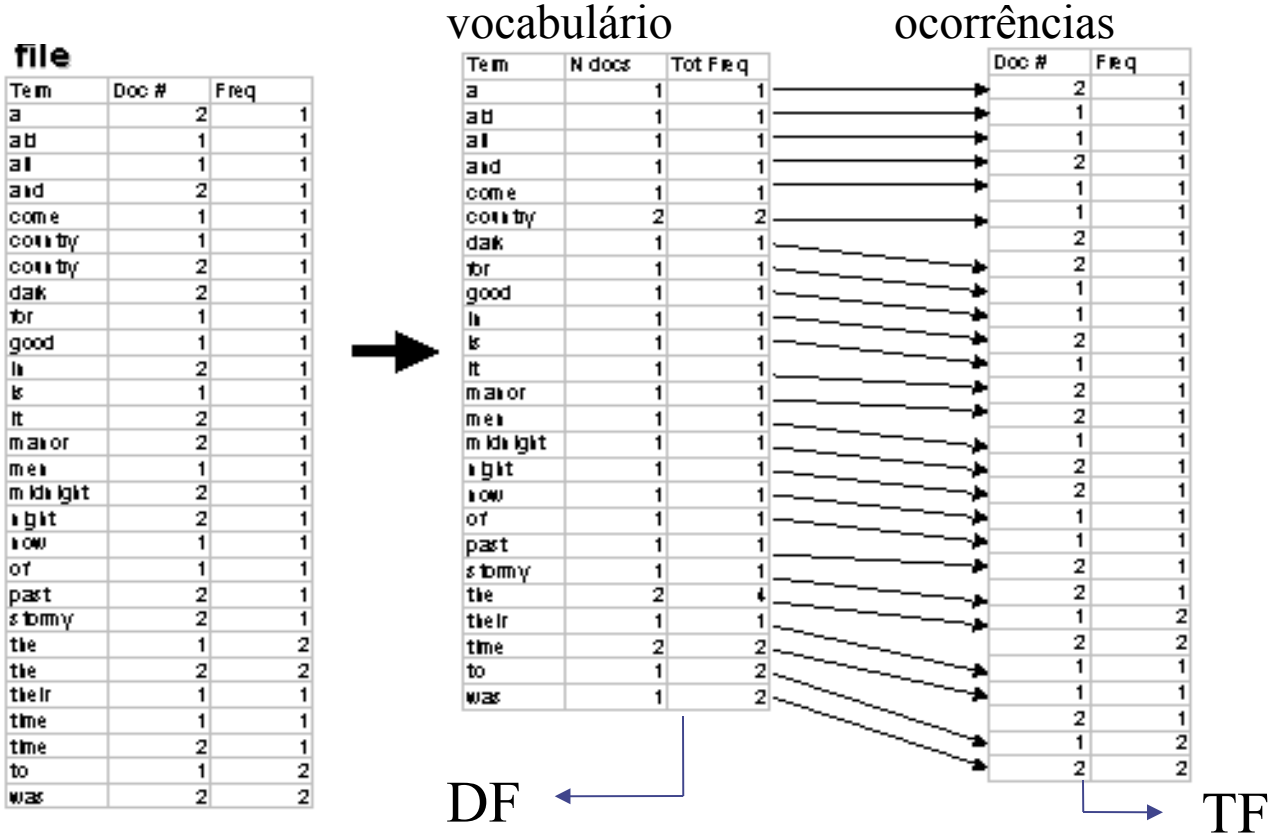
$Tf_{ij}$

# Arquivo Invertido com TF-IDF
## Construção

- A busca em um arquivo invertido sempre começa a partir do vocabulário
  - Assim, é sempre melhor armazenar o vocabulário em um arquivo separado

# Arquivo Invertido com TF-IDF
## Construção

**4.** O arquivo é então separado em duas partes: vocabulário e ocorrências




| Term | Doc # | Freq |
|---|---|---|
| a | 2 | 1 |
| abl | 1 | 1 |
| all | 1 | 1 |
| and | 2 | 1 |
| come | 1 | 1 |
| country | 1 | 1 |
| country | 2 | 1 |
| dark | 2 | 1 |
| for | 1 | 1 |
| good | 1 | 1 |
| is | 2 | 1 |
| is | 1 | 1 |
| it | 2 | 1 |
| manor | 2 | 1 |
| men | 1 | 1 |
| midnight | 2 | 1 |
| night | 2 | 1 |
| now | 1 | 1 |
| of | 1 | 1 |
| past | 2 | 1 |
| stormy | 2 | 1 |
| the | 1 | 2 |
| the | 2 | 2 |
| their | 1 | 1 |
| time | 1 | 1 |
| time | 2 | 1 |
| to | 1 | 2 |
| was | 2 | 2 |

| Term | N docs | Tot Freq |
|---|---|---|
| a | 1 | 1 |
| abl | 1 | 1 |
| all | 1 | 1 |
| and | 1 | 1 |
| come | 1 | 1 |
| country | 2 | 2 |
| dark | 1 | 1 |
| for | 1 | 1 |
| good | 1 | 1 |
| is | 1 | 1 |
| is | 1 | 1 |
| it | 1 | 1 |
| manor | 1 | 1 |
| men | 1 | 1 |
| midnight | 1 | 1 |
| night | 1 | 1 |
| now | 1 | 1 |
| of | 1 | 1 |
| past | 1 | 1 |
| stormy | 1 | 1 |
| the | 2 | 4 |
| their | 1 | 1 |
| time | 2 | 2 |
| to | 1 | 2 |
| was | 1 | 2 |

| Doc # | Freq |
|---|---|
| 2 | 1 |
| 1 | 1 |
| 1 | 1 |
| 2 | 1 |
| 1 | 1 |
| 1 | 1 |
| 2 | 1 |
| 2 | 1 |
| 1 | 1 |
| 1 | 1 |
| 2 | 1 |
| 1 | 1 |
| 2 | 1 |
| 2 | 1 |
| 1 | 1 |
| 2 | 1 |
| 2 | 1 |
| 1 | 1 |
| 1 | 1 |
| 2 | 1 |
| 2 | 1 |
| 1 | 2 |
| 2 | 2 |
| 1 | 1 |
| 1 | 1 |
| 2 | 1 |
| 1 | 2 |
| 2 | 2 |

# Arquivos Invertidos
## Busca

- O algoritmo básico segue três passos:
  - Busca do vocabulário
    - As palavras ou padrões presentes na consulta são pesquisados no vocabulário do arquivo
  - Recuperação de ocorrências
    - A lista de ocorrências de todas as palavras ou termos encontrados é recuperada
  - Manipulação de ocorrências
    - As ocorrências são processadas para resolver a consulta

# Arquivos Invertidos
## Busca

- As estruturas mais usadas para armazenar o vocabulário são tabelas *hash*, árvores e árvores-B
- A alternativa mais simples é armazenar as palavras em ordem alfabética e fazer pesquisa binária
  - Gasta menos espaço
  - Custo de tempo da ordem de *O(log n)*
    - *n* = tamanho do vocabulário

# Arquivos Invertidos
## Consultas Simples

- Consulta com apenas uma palavra
  - a busca simplesmente retorna a lista de ocorrências da palavra
  - que será utilizada na recuperação e ordenação dos documentos
- Consultas de contexto são um pouco mais complexas...

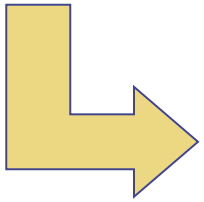
# Arquivos Invertidos
## Consultas com Contexto - Grupos Nominais

- Para consultas com GNs, o arquivo invertido deve armazenar as posições de cada palavra nos documentos
- Processo
  - Para cada palavra na consulta
    - Recupera os Doc# (identificadores) dos documentos que contêm essa palavra, e as posições onde ela ocorre
      - (Doc#;  pos1, pos2, pos3,...)
  - Faz a interseção entre os Doc# recuperados
    - Queremos os docs que contenham todas as palavras da consulta – o GN
  - Verifica a ocorrência dos GN da consulta
    - Pela posição das palavras

# Arquivos Invertidos

## Consultas com Contexto - Grupos Nominais

| Documento | Texto |
|---|---|
| 1 | Please porridge hot, pease porridge cold |
| 2 | Pease porridge in the pot |
| 3 | Nine days cold |
| 4 | Some like it hot, some like it cold |
| 5 | Some like it in the pot |
| 6 | Nine days old |



**Arquivo Invertido com posições dos termos**

| No | Termo | (Docs; Pos) |
|---|---|---|
| 1 | cold | (1;6), (4;8) |
| 2 | days | (3;2), (6;2) |
| 3 | hot | (1;3), (4;4) |
| 4 | in | (2;3), (5;4) |
| 5 | it | (4;3,7), (5;3) |
| 6 | like | (4;2,6), (5;2) |
| 7 | nine | (3;1), (6;1) |
| 8 | old | (3;3), (6;3) |
| 9 | pease | (1;1,4), (2;1) |
| 10 | porridge | (1;2,5), (2;2) |
| 11 | pot | (2;5), (5;6) |
| 12 | some | (4;1,5), (5;1) |
| 13 | the | (2;4), (5;5) |

Vocabulário

Ocorrências e posições

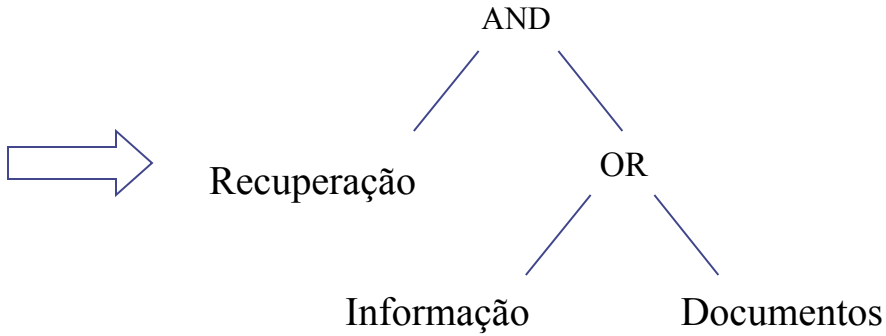
# Arquivos Invertidos
## Consultas com Contexto

- Busca com Proximidade das Palavras
- Usa uma abordagem semelhante à busca por grupos nominais
  - Seleciona os documentos em que todas as palavras da consulta ocorrem
  - Em um contexto que satisfaz as restrições de proximidade da consulta
- Exemplo de consulta: (p1, p2, 4)
  - Depois de localizar p1
  - Encontra a ocorrência mais próxima de p2 a p1
  - E verifica se está dentro da distância máxima permitida - 4

# Arquivos Invertidos
## Consultas Booleanas

- Palavras combinadas com operadores booleanos
- Cada consulta define uma árvore sintática:
  - Folhas são termos simples isolados
  - Nós internos são operadores booleanos

Consulta: Recuperação AND (Informação OR Documentos)



# Arquivos Invertidos
## Consultas Booleanas

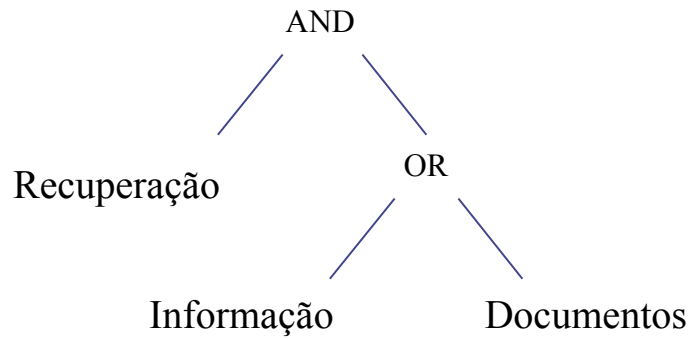
- O algoritmo de busca percorre a árvore sintática da consulta a partir das folhas
  - Folhas correspondem a buscas por palavras isoladas no arquivo invertido
  - Nós internos definem operadores sobre os conjuntos de documentos recuperados

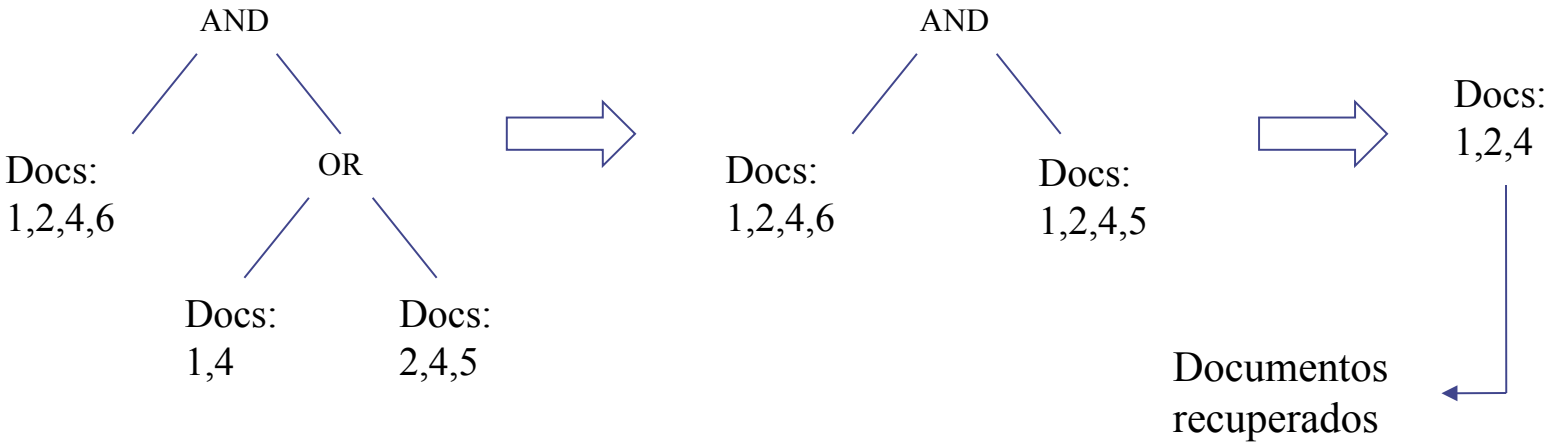
# Arquivos Invertidos
## Consultas Booleanas

- Palavra isolada
  - Recupera documentos contendo essa palavra
- OR
  - Recursivamente recupera $e_1$ e $e_2$, e faz a união dos resultados
- AND
  - Recursivamente recupera $e_1$ e $e_2$, e faz a interseção dos resultados

# Arquivos Invertidos
## Consultas Booleanas

Consulta: Recuperação AND (Informação OR Documentos)



# Exemplo de Inversões

**Exemplo de Inversão de Arquivos**

**O Arquivo de dados é exibido abaixo**

**Pede-se apresentar as inversões pelos atributos Departamento e Profissão**

| Registro | Numero | Nome | Profissão | Departamento | Arquivo | Endereço | Slot |
|---|---|---|---|---|---|---|---|
| 1 | 1000 | Ademar | Motorista | Administrativo | 1 | 6 | 1 |
| 2 | 1050 | Afonso | Programador | Técnico | 1 | 0 | 1 |
| 3 | 2400 | Iara | Secretária | Comercial | 1 | 6 | 2 |
| 4 | 1850 | Edmundo | Escriturário | Administrativo | 1 | 2 | 1 |
| 5 | 1440 | Cristiano | Diretor | Administrativo | 1 | 5 | 1 |
| 6 | 3150 | Tatiana | Diretor | Técnico | 1 | 0 | 2 |
| 7 | 2000 | Gerson | Contador | Administrativo | 1 | 5 | 2 |
| 8 | 1900 | Ênio | Almoxarife | Administrativo | 1 | 3 | 1 |
| 9 | 2430 | Ivan | Operador | Técnico | 1 | 1 | 1 |
| 10 | 2600 | Miguel | Médico | Administrativo | 1 | 3 | 2 |
| 11 | 1075 | Angela | Vendedora | Comercial | 1 | 4 | 1 |
| 12 | 1400 | Claudia | Engenheiro | Técnico | 1 | 0 | 3 |
| 13 | 2200 | Helena | Engenheiro | Técnico | 1 | 2 | 2 |
| 14 | 2700 | Ramon | Desenhista | Técnico | 1 | 5 | 3 |
| 15 | 1100 | Antônio | Gerente | Técnico | 1 | 1 | 2 |
| 16 | 1800 | Edson | Escriturário | Comercial | 1 | 1 | 3 |
| 17 | 3100 | Sônia | Vendedora | Comercial | 1 | 6 | 3 |
| 18 | 2500 | Maria | Secretária | Técnico | 1 | 1 | 4 |
| 19 | 1300 | Carlos | Publicitário | Comercial | 1 | 5 | 4 |
| 20 | 2450 | Luiz | Publicitário | Comercial | 1 | 0 | 4 |
| 21 | 1600 | Diogo | Contínuo | Administrativo | 1 | 4 | 2 |
| 22 | 1480 | Darci | Contador | Administrativo | 1 | 3 | 3 |
| 23 | 2950 | Sandra | Analista | Técnico | 1 | 3 | 4 |
| 24 | 1970 | Genaro | Engenheiro | Técnico | 1 | 3 | 5 |
| 25 | 1350 | César | Gerente | Comercial | 1 | 6 | 4 |
| 26 | 1950 | Flávio | Gerente | Administrativo | 1 | 4 | 3 |
| 27 | 1700 | Éber | Analista | Técnico | 1 | 6 | 5 |

# Inversão por Departamento usando encadeamento

**Exemplo de Inversão de Arquivos**

**Inversão por Departamento**

| Registro | Arquivo | Endereço | Slot | Nome | Departamento | Próx. | Anter. |
|---|---|---|---|---|---|---|---|
| 0 | | | | | Caixa de nós | 31 | 31 |
| 1 | 9 | | | | Administrativo | 4 | 29 |
| 2 | 11 | | | | Técnico | 5 | 30 |
| 3 | 7 | | | | Comercial | 6 | 28 |
| | | | | | | | |
| 4 | 1 | 6 | 1 | Ademar | Administrativo | 7 | 1 |
| 5 | 1 | 0 | 1 | Afonso | Técnico | 9 | 2 |
| 6 | 1 | 6 | 2 | Iara | Comercial | 14 | 3 |
| 7 | 1 | 2 | 1 | Edmundo | Administrativo | 8 | 4 |
| 8 | 1 | 5 | 1 | Cristiano | Administrativo | 10 | 7 |
| 9 | 1 | 0 | 2 | Tatiana | Técnico | 12 | 5 |
| 10 | 1 | 5 | 2 | Gerson | Administrativo | 11 | 8 |
| 11 | 1 | 3 | 1 | Ênio | Administrativo | 13 | 10 |
| 12 | 1 | 1 | 1 | Ivan | Técnico | 15 | 9 |
| 13 | 1 | 3 | 2 | Miguel | Administrativo | 24 | 11 |
| 14 | 1 | 4 | 1 | Angela | Comercial | 19 | 6 |
| 15 | 1 | 0 | 3 | Claudia | Técnico | 16 | 12 |
| 16 | 1 | 2 | 2 | Helena | Técnico | 17 | 15 |
| 17 | 1 | 5 | 3 | Ramon | Técnico | 18 | 16 |
| 18 | 1 | 1 | 2 | Antônio | Técnico | 21 | 17 |
| 19 | 1 | 1 | 3 | Edson | Comercial | 20 | 14 |
| 20 | 1 | 6 | 3 | Sônia | Comercial | 22 | 19 |
| 21 | 1 | 1 | 4 | Maria | Técnico | 26 | 18 |
| 22 | 1 | 5 | 4 | Carlos | Comercial | 23 | 20 |
| 23 | 1 | 0 | 4 | Luiz | Comercial | 28 | 22 |
| 24 | 1 | 4 | 2 | Diogo | Administrativo | 25 | 13 |
| 25 | 1 | 3 | 3 | Darci | Administrativo | 29 | 24 |
| 26 | 1 | 3 | 4 | Sandra | Técnico | 27 | 21 |
| 27 | 1 | 3 | 5 | Genaro | Técnico | 30 | 26 |
| 28 | 1 | 6 | 4 | César | Comercial | 3 | 23 |
| 29 | 1 | 4 | 3 | Flávio | Administrativo | 1 | 25 |
| 30 | 1 | 6 | 5 | Éber | Técnico | 2 | 27 |

# Inversão por Profissão usando encadeamento

**Exemplo de Inversão de Arquivos**

**Inversão por Profissão**

| Registro | Arquivo | Endereço | Slot | Nome | Profissão | Próx. | Anter. |
|---|---|---|---|---|---|---|---|
| 0 | | | | | Caixa de nós | 44 | 44 |
| 1 | 1 | | | | Motorista | 35 | 35 |
| 2 | 1 | | | | Programador | 37 | 37 |
| 3 | 2 | | | | Secretária | 40 | 41 |
| 4 | 2 | | | | Escriturário | 29 | 30 |
| 5 | 2 | | | | Diretor | 24 | 25 |
| 6 | 2 | | | | Contador | 20 | 21 |
| 7 | 1 | | | | Almoxarife | 17 | 17 |
| 8 | 1 | | | | Operador | 36 | 36 |
| 9 | 1 | | | | Médico | 34 | 34 |
| 10 | 2 | | | | Vendedora | 42 | 43 |
| 11 | 3 | | | | Engenheiro | 26 | 28 |
| 12 | 1 | | | | Desenhista | 23 | 23 |
| 13 | 3 | | | | Gerente | 31 | 33 |
| 14 | 2 | | | | Publicitário | 38 | 39 |
| 15 | 1 | | | | Contínuo | 22 | 22 |
| 16 | 2 | | | | Analista | 18 | 19 |
| | | | | | | | |
| 17 | 1 | 6 | 1 | Ademar | Motorista | 1 | 1 |
| 18 | 1 | 0 | 1 | Afonso | Programador | 2 | 2 |
| 19 | 1 | 6 | 2 | Iara | Secretaria | 34 | 3 |
| 20 | 1 | 2 | 1 | Edmundo | Escriturário | 32 | 4 |
| 21 | 1 | 5 | 1 | Cristiano | Diretor | 22 | 5 |
| 22 | 1 | 0 | 2 | Tatiana | Diretor | 5 | 21 |
| 23 | 1 | 5 | 2 | Gerson | Contador | 38 | 6 |
| 24 | 1 | 3 | 1 | Ênio | Almoxarife | 7 | 7 |
| 25 | 1 | 1 | 1 | Ivan | Operador | 8 | 8 |
| 26 | 1 | 3 | 2 | Miguel | Médico | 9 | 9 |
| 27 | 1 | 4 | 1 | Angela | Vendedora | 33 | 10 |
| 28 | 1 | 0 | 3 | Claudia | Engenheiro | 29 | 11 |
| 29 | 1 | 2 | 2 | Helena | Engenheiro | 40 | 28 |
| 30 | 1 | 5 | 3 | Ramon | Desenhista | 12 | 12 |
| 31 | 1 | 1 | 2 | Antônio | Gerente | 41 | 13 |
| 32 | 1 | 1 | 3 | Edson | Escriturário | 4 | 20 |
| 33 | 1 | 6 | 3 | Sônia | Vendedora | 10 | 27 |
| 34 | 1 | 1 | 4 | Maria | Secretária | 3 | 19 |
| 35 | 1 | 5 | 4 | Carlos | Publicitário | 36 | 14 |

# Arquivos de Assinaturas

# Arquivos de Assinaturas

- Uma alternativa aos arquivos de índices invertidos
  - Ganha na velocidade de busca/recuperação de documentos

# Arquivos de Assinaturas

- Estrutura de indexação baseada em vetores binários
  - Cada palavra no vocabulário da base de documentos é mapeada em um vetor de B-bits
    - Sua assinatura
  - B é fixo e depende do tamanho do vocabulário da base de documentos
  - O mapeamento é feito através de funções de *hash,* com duas possibilidades:
    - Uma função única que define os valores de todos os bits de uma vez, ou
    - Uma função diferente para definir cada bit do vetor

# Arquivos de Assinaturas
## Vocabulário da Base de Documentos

- Os vetores das assinaturas raramente coincidem
  - para vetores com um tamanho adequado ao tamanho do vocabulário
  - Para boas funções de *hash*
- Porém, os valores dos bits na vertical podem coincidir
  - Problemas de precisão na recuperação

| Termos | Assinaturas com 16 bits |
|---|---|
| cold | 1000 0000 0010 0100 |
| days | 0010 0100 0000 1000 |
| hot | 0000 1010 0000 0000 |
| in | 0000 1001 0010 0000 |
| it | 0000 1000 1000 0010 |
| like | 0100 0010 0000 0001 |
| nine | 0010 1000 0000 0100 |
| old | 1000 1000 0100 0000 |
| pease | 0000 0101 0000 0001 |
| porridge | 0100 0100 0010 0000 |
| pot | 0000 0010 0110 0000 |
| some | 0100 0100 0000 0001 |
| the | 1010 1000 0000 0000 |

# Arquivos de Assinaturas
## Assinatura dos Documentos

- A assinatura de cada documento pode ser obtida com base nas assinaturas das suas palavras
  - Aplicando o operador OR às assinaturas dos termos que aparecem no documento

| Documento | Texto | Assinatura |
|---|---|---|
| 1 | Pease porridge hot, pease porridge cold, | 1100 1111 0010 0101 |
| 2 | Pease porridge in the pot, | 1110 1111 0110 0001 |
| 3 | Nine days old. | 1010 1100 0100 1100 |
| 4 | Some like it hot, some like it cold. | 1100 1110 1010 0111 |
| 5 | Some like it in the pot, | 1110 1111 1110 0011 |
| 6 | Nine days old. | 1010 1100 0100 1100 |

# Arquivos de Assinaturas
## Consultas

- Procedimento para consultas com uma palavra
  - A palavra é mapeada na sua assinatura com as mesmas funções utilizadas no mapeamento do vocabulário da base
  - Realiza-se uma busca seqüencial na base de assinaturas dos documentos procurando por documentos relevantes
    - Usando o operador AND para comparar os vetores

# Arquivos de Assinaturas
## Consultas

- Formalização:
  - Seja Bj a assinatura do documento Dj
  - Seja P a assinatura da palavra da consulta
  - Então recupere todos os documentos em que
    P AND Bj = P
    - Esses documentos provavelmente contêm a palavra da consulta

# Arquivos de Assinaturas
## Consultas

- Em outras palavras...
  - Se qualquer bit com valor = 1 na assinatura da consulta tiver valor = 0 na assinatura do documento, então com certeza o documento não contém a palavra da consulta
  - Se todos os bits = 1 da assinatura da consulta também têm valor = 1 no documento, então provavelmente a palavra da consulta está presente no documento
    - Por que "provavelmente" ?

# Arquivos de Assinaturas
# Dificuldades

- É possível que
  - todos os bits =1 na assinatura da consulta tenham valor = 1 no documento também
  - mas o termo não esteja presente no documento (*false drop*)
- Probabilidade de *false drop* é maior para documentos com muitos termos
  - uma vez que teriam assinatura com muitos bits iguais a 1
- Aumentando o tamanho da assinatura, diminuímos a probabilidade de *false drop*

Exemplo:

pesquisar por “like” AND “pot”

LIKE = 0100 0010 0000 0001
POT = 0000 0010 0110 0000
L&P = 0100 0010 0110 0001

L&P = like AND pot

Operação realizada: OR

| Termos | Assinaturas com 16 bits |
|---|---|
| cold | 1000 0000 0010 0100 |
| days | 0010 0100 0000 1000 |
| hot | 0000 1010 0000 0000 |
| in | 0000 1001 0010 0000 |
| it | 0000 1000 1000 0010 |
| like | 0100 0010 0000 0001 |
| nine | 0010 1000 0000 0100 |
| old | 1000 1000 0100 0000 |
| pease | 0000 0101 0000 0001 |
| porridge | 0100 0100 0010 0000 |
| pot | 0000 0010 0110 0000 |
| some | 0100 0100 0000 0001 |
| the | 1010 1000 0000 0000 |

| Doc | Texto | Assinatura |
|---|---|---|
| 1 | Pease porridge hot, pease porridge cold, | 1100 1111 0010 0101 |
| 2 | Pease porridge in the pot, | 1110 1111 0110 0001 |
| 3 | Nine days old. | 1010 1100 0100 1100 |
| 4 | Some like it hot, some like it cold, | 1100 1110 1010 0111 |
| 5 | Some like it in the pot, | 1110 1111 1110 0011 |
| 6 | Nine days old. | 1010 1100 0100 1100 |

Pesquisando pela assinatura 0100 0010 0110 0001 na base de documentos, teremos como resposta os documentos 2 e 5. Deveria retornar somente o documento 5:

1 1100 1111 0010 0101
0100 0010 0110 0001
0100 0010 0010 0001

2 1110 1111 0110 0001
0100 0010 0110 0001
0100 0010 0110 0001 OK

3 1010 1100 0100 1100
0100 0010 0110 0001
0000 0000 0100 0000

4 1100 1110 1010 0111
0100 0010 0110 0001
0100 0010 0010 0001

5 1110 1111 1110 0011
0100 0010 0110 0001
0100 0010 0110 0001 OK

6 1010 1100 0100 1100
0000 0111 0110 0001
0000 0100 0100 0000

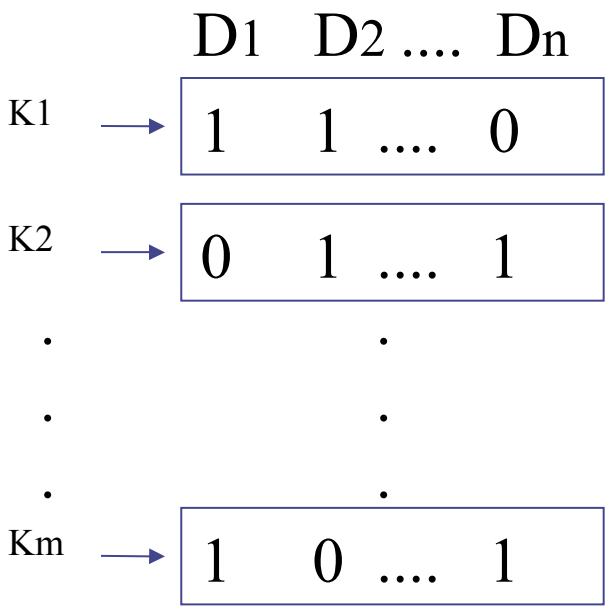
# Bitmaps

# Bitmaps

- Estrutura que também trabalha com valores binários, porém utiliza um procedimento diferente para criar as assinaturas
- Cria uma matriz de termos (Ki) x documentos (Dj) da base
  - Se o termo Ki está presente no documento Dj, então o elemento ij da matriz é =1
  - caso contrário, ij=0

# Bitmaps - Exemplo

- Conjunto de n documentos indexados através de m termos

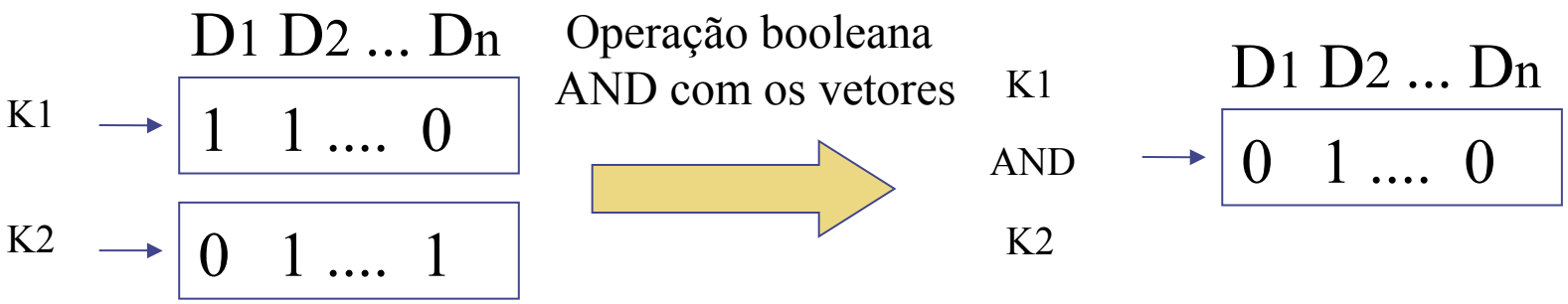


# Bitmaps
## Consultas

- Para consultas com um termo simples
  - pesquisa o vetor do termo (linha da matriz) de forma seqüencial
    - Compara bit a bit
  - retorna os documentos com valor do bit=1
- Consultas booleanas também são simples
  - Recupera as linhas dos termos da consulta
  - Aplica o operador booleano da consulta
  - Só depois faz a pesquisa seqüencial bit a bit

# Bitmaps – Exemplo de Consulta

- Considere a consulta Q = K1 AND K2




- Uma pesquisa seqüencial no vetor K1 AND K2 irá retornar os documentos que satisfazem a consulta

# Bitmaps

- Método ocupa muito espaço desnecessário para termos pouco comuns
  - Maioria dos bits iguais a 0
- É ineficiente para adicionar e deletar documentos
  - Uma vez que se deve verificar a presença ou ausência de todos os termos no documento
  - Nos arquivos invertidos, trabalha-se apenas com os termos que aparecem de fato no documento

# Inversão por Profissão usando Mapas de Bits

**Exemplo de Inversão de Arquivos**

**Inversão por Profissão usando mapas de bits**

| Regist | Motorista | Programador | Secretária | Escriturário | Diretor | Contador | Almoxarife | Operador | Médico | Vendedora | Engenheiro | Desenhista | Gerente | Publicitário | Contínuo | Analista |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 16 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 19 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 21 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 22 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 23 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 24 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 26 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |

# Conclusões

- Arquivos invertidos são os mais usados em sistemas de Recuperação de Informação
  - uma vez que podem ser usados para resolver uma grande quantidade de tipos de consultas
- Arquivos de assinaturas e Bitmaps são usados basicamente para consultas com termos simples e consultas booleanas
- Arquivo de assinaturas é muito estudado, mas pouco usado