



Universidade Federal do Rio de Janeiro (UFRJ)
Departamento de Ciência da Computação (DCC)



Recuperação da Informação (MAB605)

Modelo Booleano

Profa. Giseli Rabello Lopes

Roteiro

- Introdução
- Modelo booleano
 - Matriz de incidência
 - Índice invertido
 - Otimização de consultas
- Referências

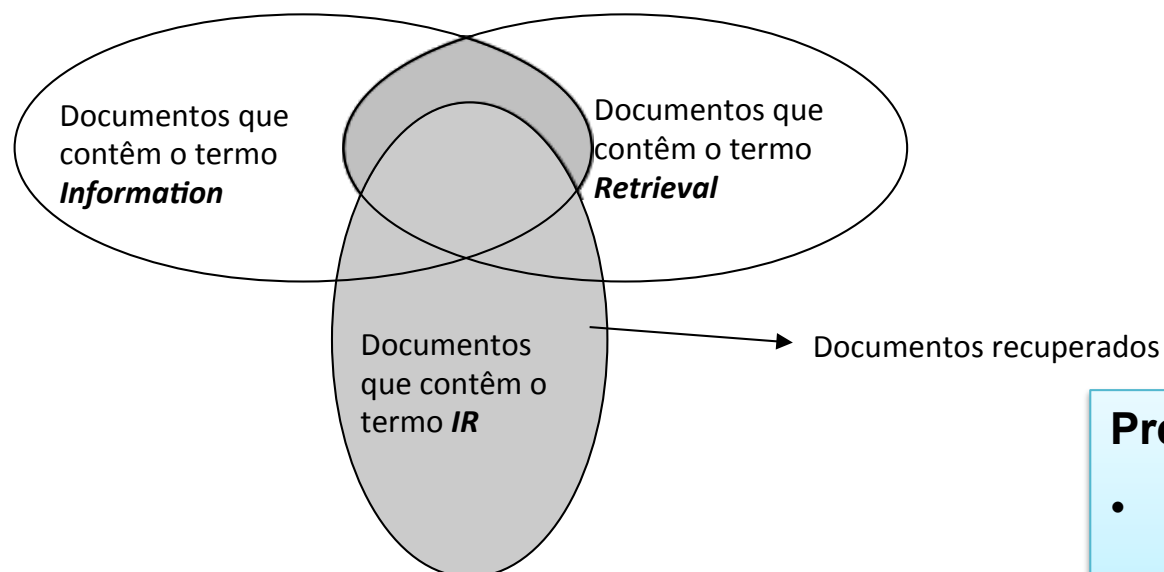
Modelo Booleano

- Modelo de RI simples baseado em:
 - Teoria dos conjuntos
 - Álgebra booleana
- Representações:
 - Documentos (***D***)
 - Conjuntos de termos de indexação
 - Consultas (***Q***)
 - Formuladas através de expressões booleanas
 - Termos e conectivos de *boole* (**AND**, **OR** e **NOT**)
 - Operações sobre conjuntos:
 - Intersecção (**AND**)
 - União (**OR**)
 - Negação (**NOT**)

Modelo Booleano

- Exemplo:

(Information AND Retrieval) OR IR



Precedência:

- NÃO (*NOT*)
- E (*AND*)
- OU (*OR*)

Modelo Booleano

- Resultado:
 - Critério de decisão binário
 - Função de similaridade:

$$sim(d_j, q) = \begin{cases} \mathbf{1} & \text{se } d_j \text{ satisfaz condições da expressão booleana } q \\ \mathbf{0} & \text{caso contrário} \end{cases}$$

Modelo Booleano

- Matriz de incidências

	doc_1	doc_2	doc_3	doc_j
$termo_1$	0	1	0	1
$termo_2$	1	1	0	1
$termo_3$	0	0	1	1
$termo_i$	0	0	0	1

– Consulta = $termo_1 \wedge \neg termo_3$
 $0101 \wedge \neg 0011$
 $0101 \wedge 1100 = 0100 (doc_2)$

Matriz de incidência termos-documentos

	Antony and Cleopatra	Julius Caesar	The Thempest	Hamlet	Othello	Macbeth	...
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	

[Manning et al., 2008]

Brutus AND Caesar AND NOT Calpurnia

110100 AND 110111 AND 101111 = 100100

Matriz de incidência termos-documentos

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	...
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	

[Manning et al., 2008]

Brutus AND Caesar AND NOT Calpurnia

110100 AND 110111 AND 101111 = 100100

Matriz de incidência termos-documentos

	Antony and Cleopatra	Julius Caesar	The Thempest	Hamlet	Othello	Macbeth	...
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	

[Manning et al., 2008]

Brutus AND Caesar AND NOT Calpurnia

110100 AND 110111 AND 101111 = 100100

Matriz de incidência termos-documentos


	Antony and Cleopatra	Julius Caesar	The Thempest	Hamlet	Othello	Macbeth	...
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	

[Manning et al., 2008]

Brutus AND Caesar AND NOT Calpurnia

110100 AND 110111 AND 101111 = 100100

Matriz de incidência termos-documentos



	Antony and Cleopatra	Julius Caesar	The Thempest	Hamlet	Othello	Macbeth	...
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	

[Manning et al., 2008]

Brutus AND Caesar AND NOT Calpurnia

110100 AND 110111 AND 101111 = 100100

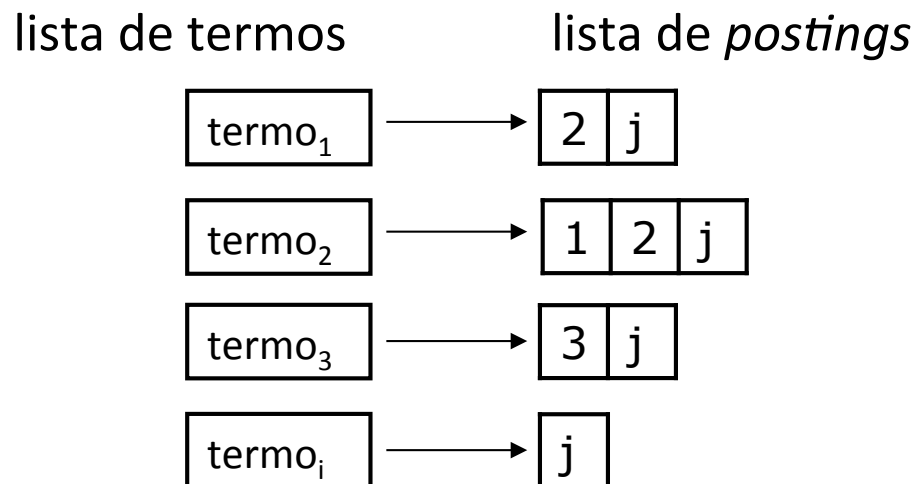
Modelo Booleano

- Matriz de incidências não é adequada para coleções de tamanho médio e grande
 - Matriz muito grande e esparsa
 - Espaço de armazenamento
 - Tempo de processamento
- Indexação
 - Arquivo invertido

Modelo Booleano

- Indexação
 - Arquivo invertido

	doc_1	doc_2	doc_3	doc_j
$termo_1$	0	1	0	1
$termo_2$	1	1	0	1
$termo_3$	0	0	1	1
$termo_i$	0	0	0	1



Modelo Booleano

- Indexação

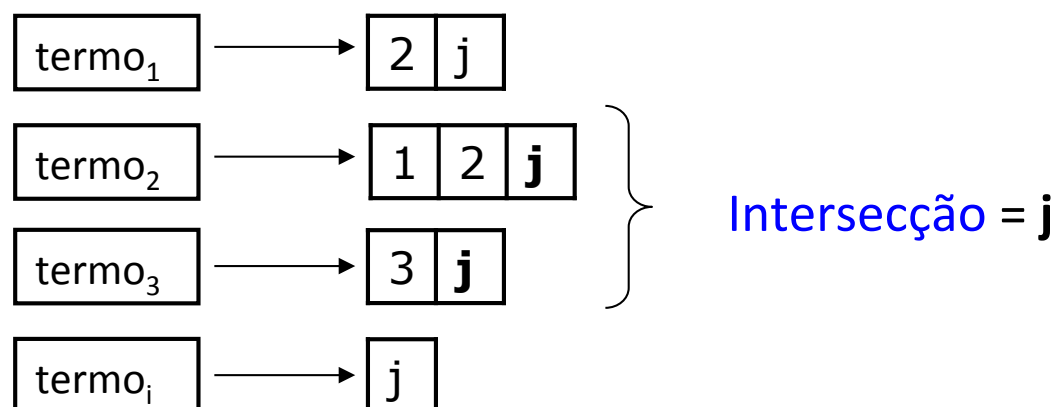
- Arquivo invertido

- Consulta = $\text{termo}_2 \wedge \text{termo}_3$

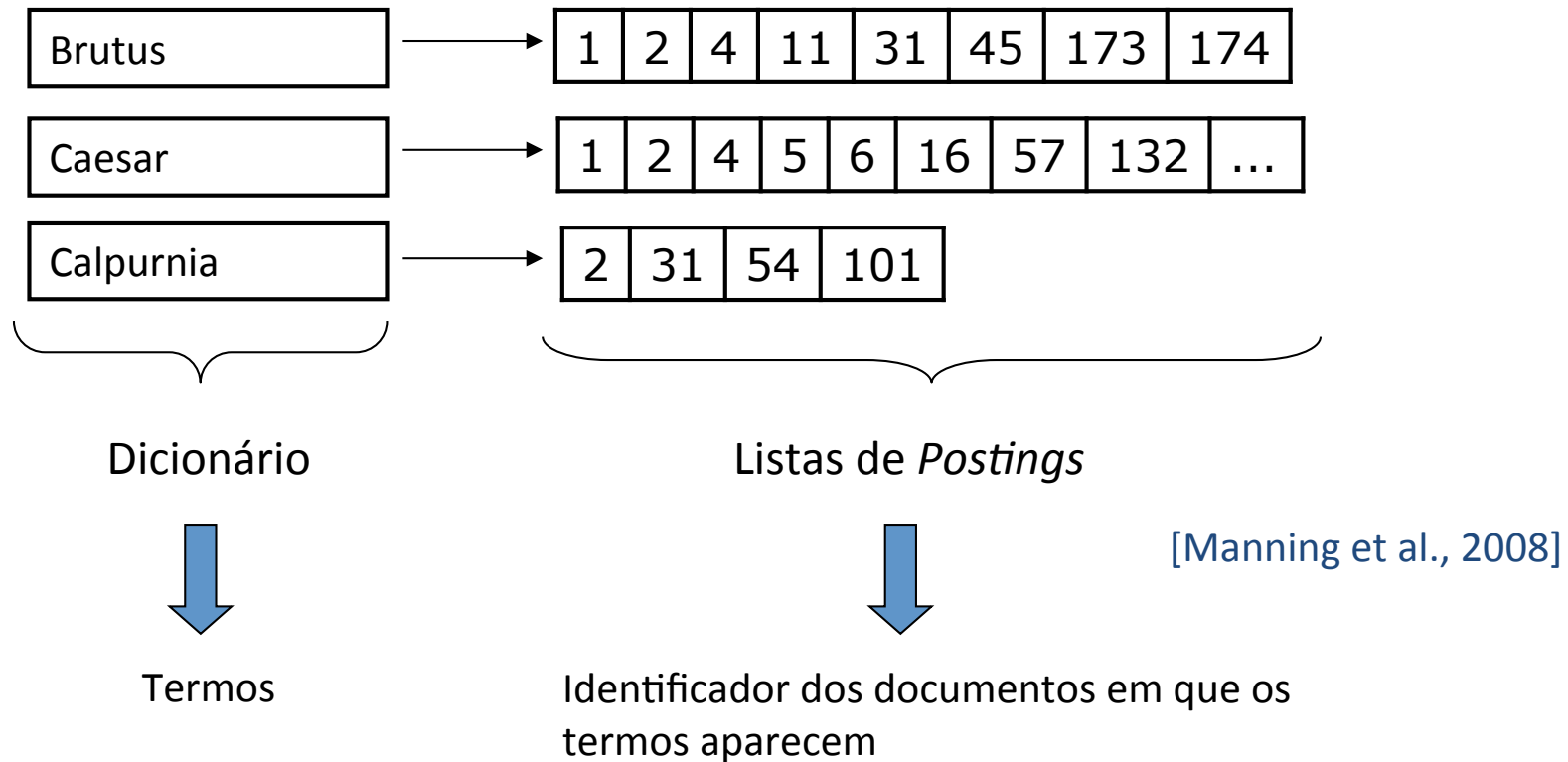
	doc_1	doc_2	doc_3	doc_j
termo_1	0	1	0	1
termo_2	1	1	0	1
termo_3	0	0	1	1
termo_i	0	0	0	1

lista de termos

lista de *postings*

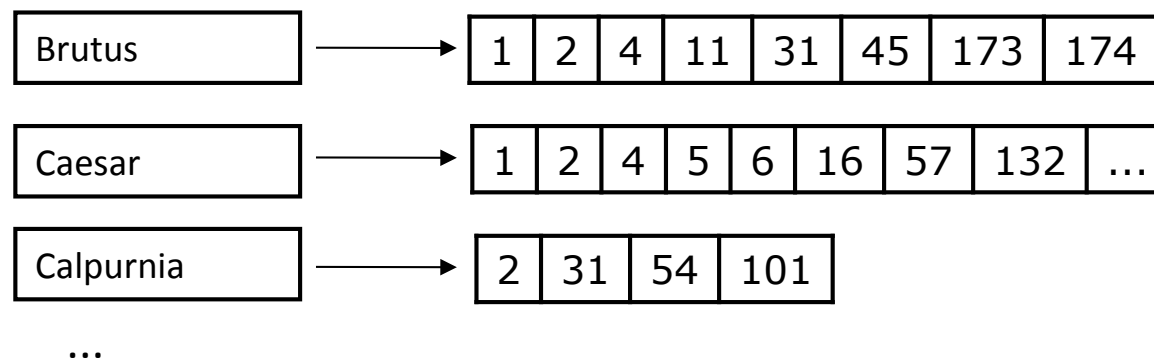


Índice invertido



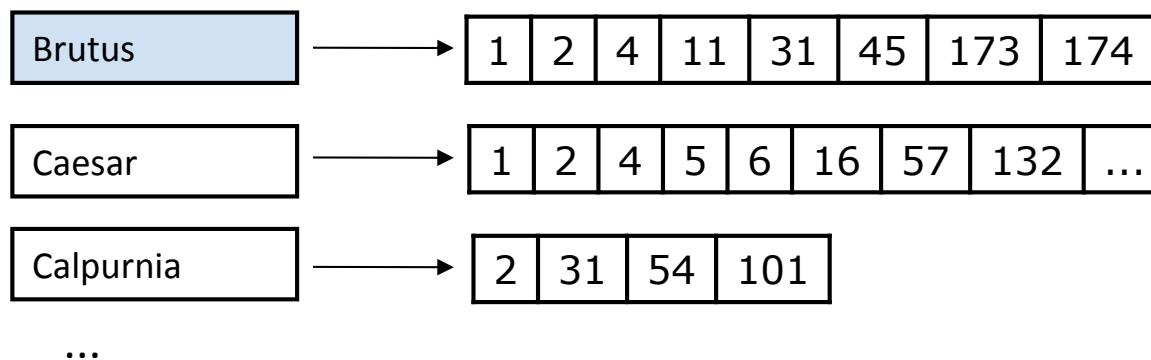
Índice invertido

- Consulta simples: **Brutus AND Calpurnia**
 - Resolução:
 1. Localizar Brutus no Dicionário;
 2. Recuperar sua lista de *postings*;
 3. Localizar Calpurnia no Dicionário;
 4. Recuperar sua lista de *postings*;
 5. Calcular a intersecção entre as duas listas de *postings*.



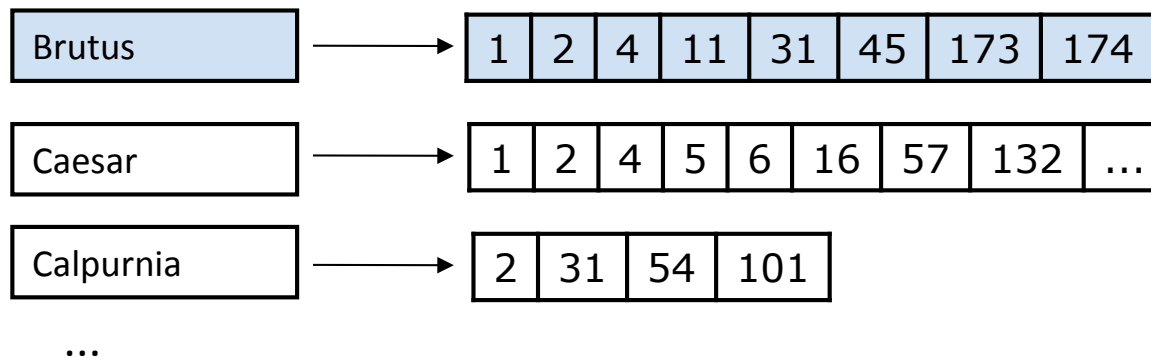
Índice invertido

- Consulta simples: **Brutus AND Calpurnia**
 - Resolução:
 1. Localizar Brutus no Dicionário;
 2. Recuperar sua lista de *postings*;
 3. Localizar Calpurnia no Dicionário;
 4. Recuperar sua lista de *postings*;
 5. Calcular a intersecção entre as duas listas de *postings*.



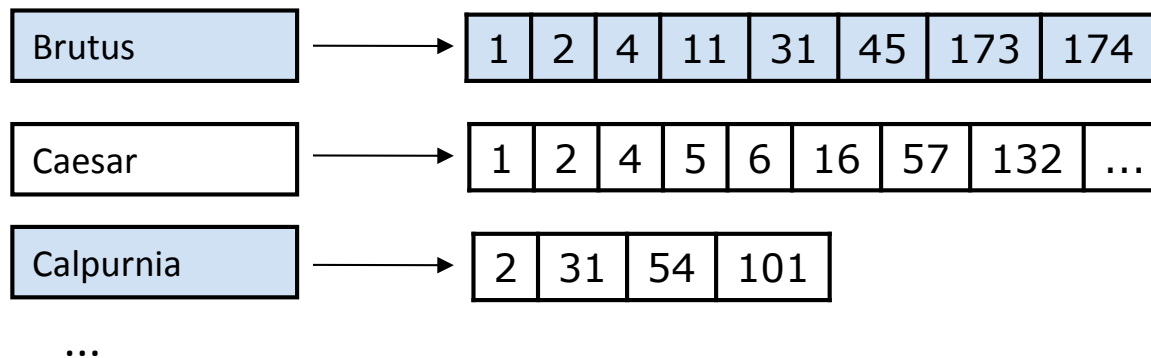
Índice invertido

- Consulta simples: **Brutus AND Calpurnia**
 - Resolução:
 1. Localizar Brutus no Dicionário;
 2. Recuperar sua lista de *postings*;
 3. Localizar Calpurnia no Dicionário;
 4. Recuperar sua lista de *postings*;
 5. Calcular a intersecção entre as duas listas de *postings*.



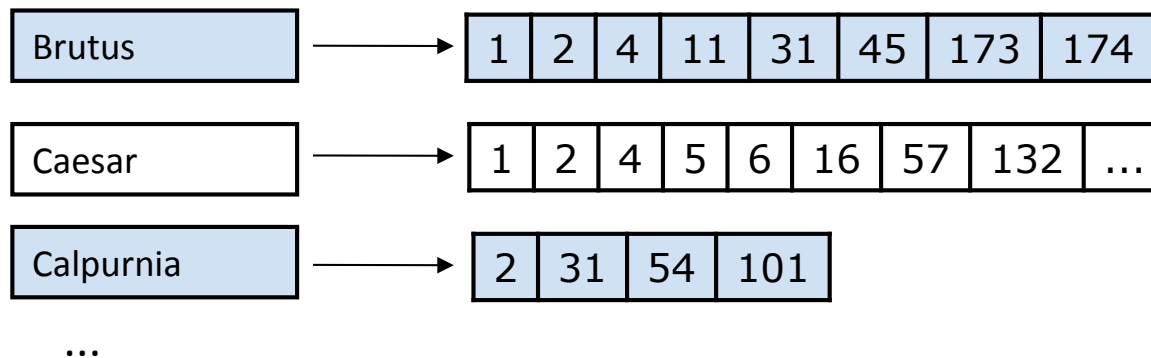
Índice invertido

- Consulta simples: **Brutus AND Calpurnia**
 - Resolução:
 1. Localizar Brutus no Dicionário;
 2. Recuperar sua lista de *postings*;
 3. Localizar Calpurnia no Dicionário;
 4. Recuperar sua lista de *postings*;
 5. Calcular a intersecção entre as duas listas de *postings*.



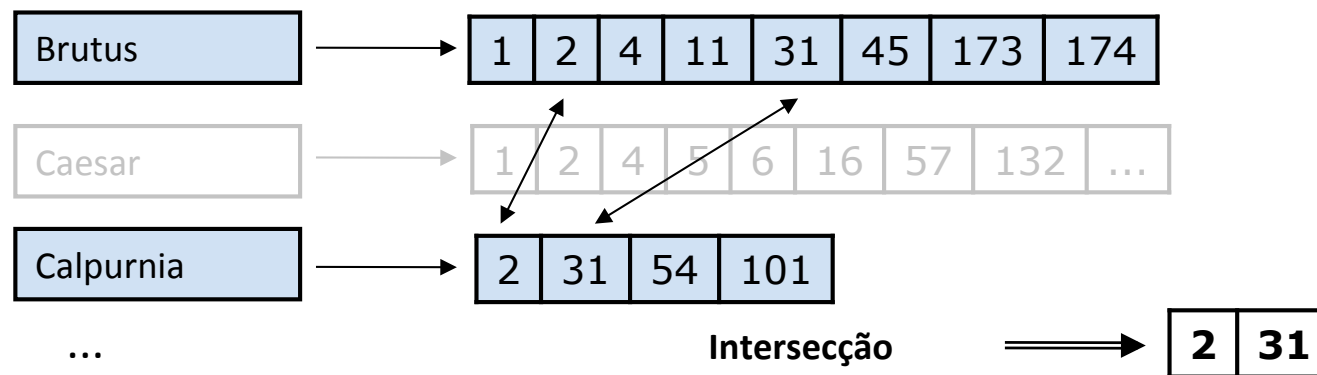
Índice invertido

- Consulta simples: **Brutus AND Calpurnia**
 - Resolução:
 1. Localizar Brutus no Dicionário;
 2. Recuperar sua lista de *postings*;
 3. Localizar Calpurnia no Dicionário;
 4. Recuperar sua lista de *postings*;
 5. Calcular a intersecção entre as duas listas de *postings*.



Índice invertido

- Consulta simples: **Brutus AND Calpurnia**
 - Resolução:
 1. Localizar Brutus no Dicionário;
 2. Recuperar sua lista de *postings*;
 3. Localizar Calpurnia no Dicionário;
 4. Recuperar sua lista de *postings*;
 5. Calcular a intersecção entre as duas listas de *postings*.



Interseção [Manning et al., 2008]

- Algoritmo para intersecção de duas listas de *postings* p_1 e p_2

```
INTERSECT( $p_1, p_2$ )  
1   $answer \leftarrow \langle \rangle$   
2  while  $p_1 \neq \text{NIL}$  and  $p_2 \neq \text{NIL}$   
3  do if  $\text{docID}(p_1) = \text{docID}(p_2)$   
4      then  $\text{ADD}(answer, \text{docID}(p_1))$   
5           $p_1 \leftarrow \text{next}(p_1)$   
6           $p_2 \leftarrow \text{next}(p_2)$   
7  else if  $\text{docID}(p_1) < \text{docID}(p_2)$   
8      then  $p_1 \leftarrow \text{next}(p_1)$   
9      else  $p_2 \leftarrow \text{next}(p_2)$   
10 return  $answer$ 
```

Obs.: Listas de *postings* devem estar ordenadas por *docID*

Exercício

- Considere uma coleção formada pelos documentos a seguir:
 - d_1 = “um navegador explorou o oceano”
 - d_2 = “mozilla firefox é o melhor navegador”
 - d_3 = “internet explorer versus firefox”
- Construa um índice invertido considerando apenas os termos presentes na consulta abaixo e simule sua execução:
 - $q = (\text{navegador AND NOT oceano}) \text{ OR internet OR firefox}$

Otimização de consultas

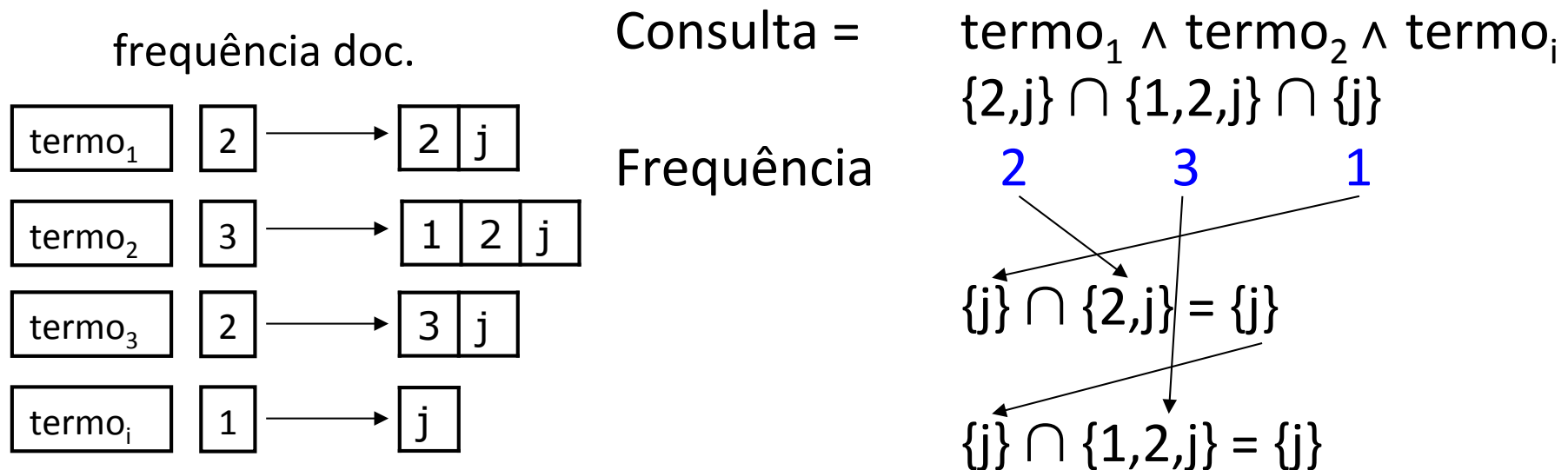
- Objetiva organizar o trabalho de resposta para que a consulta seja realizada no menor tempo e com o menor trabalho possível
 - Tentar executar as operações numa ordem que vise que menores resultados sejam gerados ao longo da execução
 - Identificar a melhor ordem de acesso nas listas de *postings*

Otimização de consultas [Manning et al., 2008]

- Estimar tamanho do resultado
 - AND (Intersecção)
 - Será no máximo igual ao tamanho da menor lista de *postings*
 - OR (União)
 - Será no máximo igual a soma do tamanho das duas listas de *postings*
 - NOT (Negação)
 - Será diferença entre número de documentos da coleção e o tamanho da lista de *postings* do termo

Otimização de consultas

- AND
 - Tamanho máximo do resultado é o menor tamanho das listas de *postings*
 - Começar com os termos de menor frequência



Consultas conjuntivas [Manning et al., 2008]

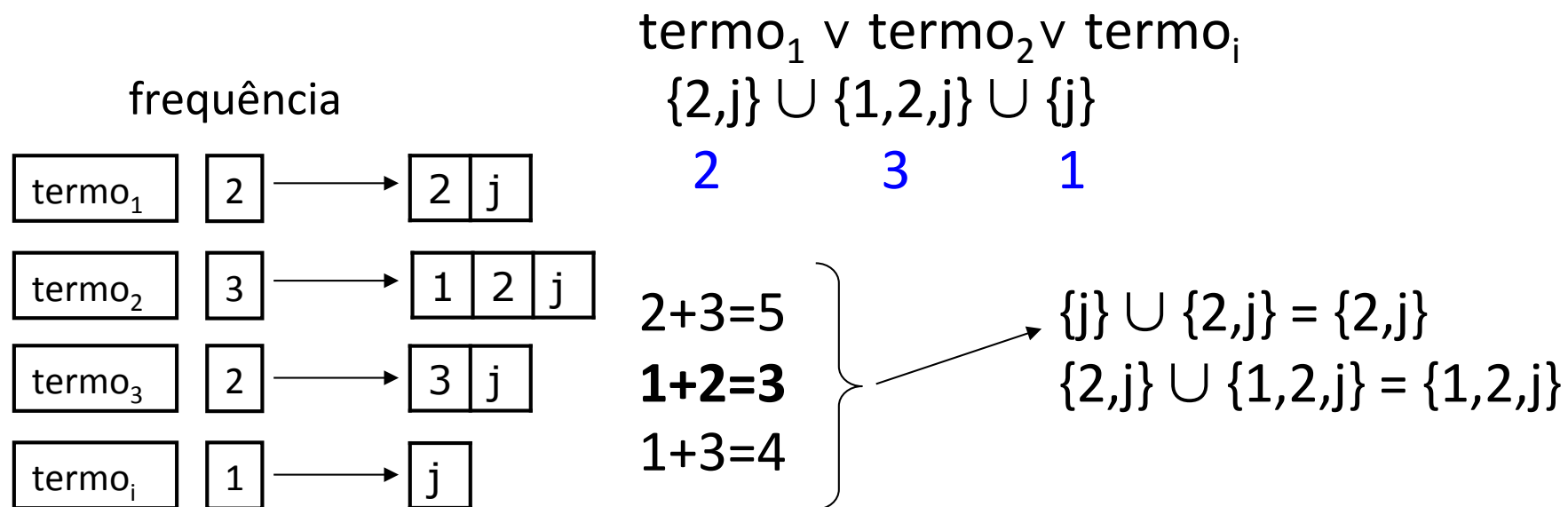
- Algoritmo para consultas conjuntivas que retorna o conjunto de documentos contendo cada termo da lista de entrada

INTERSECT($\langle t_1, \dots, t_n \rangle$)

```
1  terms  $\leftarrow$  SORTBYINCREASINGFREQUENCY( $\langle t_1, \dots, t_n \rangle$ )
2  result  $\leftarrow$  postings(first(terms))
3  terms  $\leftarrow$  rest(terms)
4  while terms  $\neq$  NIL and result  $\neq$  NIL
5  do result  $\leftarrow$  INTERSECT(result, postings(first(terms)))
6      terms  $\leftarrow$  rest(terms)
7  return result
```

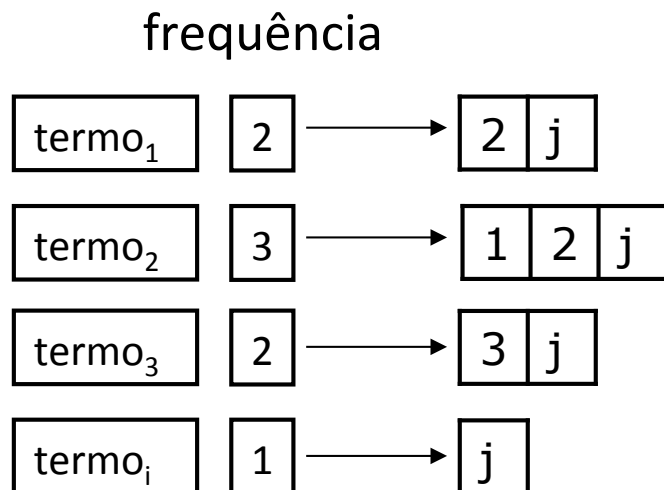
Otimização de consultas

- OR
 - Tamanho máximo do resultado é a soma dos tamanhos das listas de *postings*
 - Começar com operações com menor estimativa



Otimização de consultas

- AND + OR
 - Utilizar heurística mista



$$(\text{termo}_1 \vee \text{termo}_2) \wedge (\text{termo}_3 \vee \text{termo}_i)$$
$$\{ \{2,j\} \cup \{1,2,j\} \} \cap \{ \{3,j\} \cup \{j\} \}$$

$$2 + 3 = 5$$

$$2 + 1 = 3$$

$$\{3,j\} \cup \{j\} = \{3,j\}$$

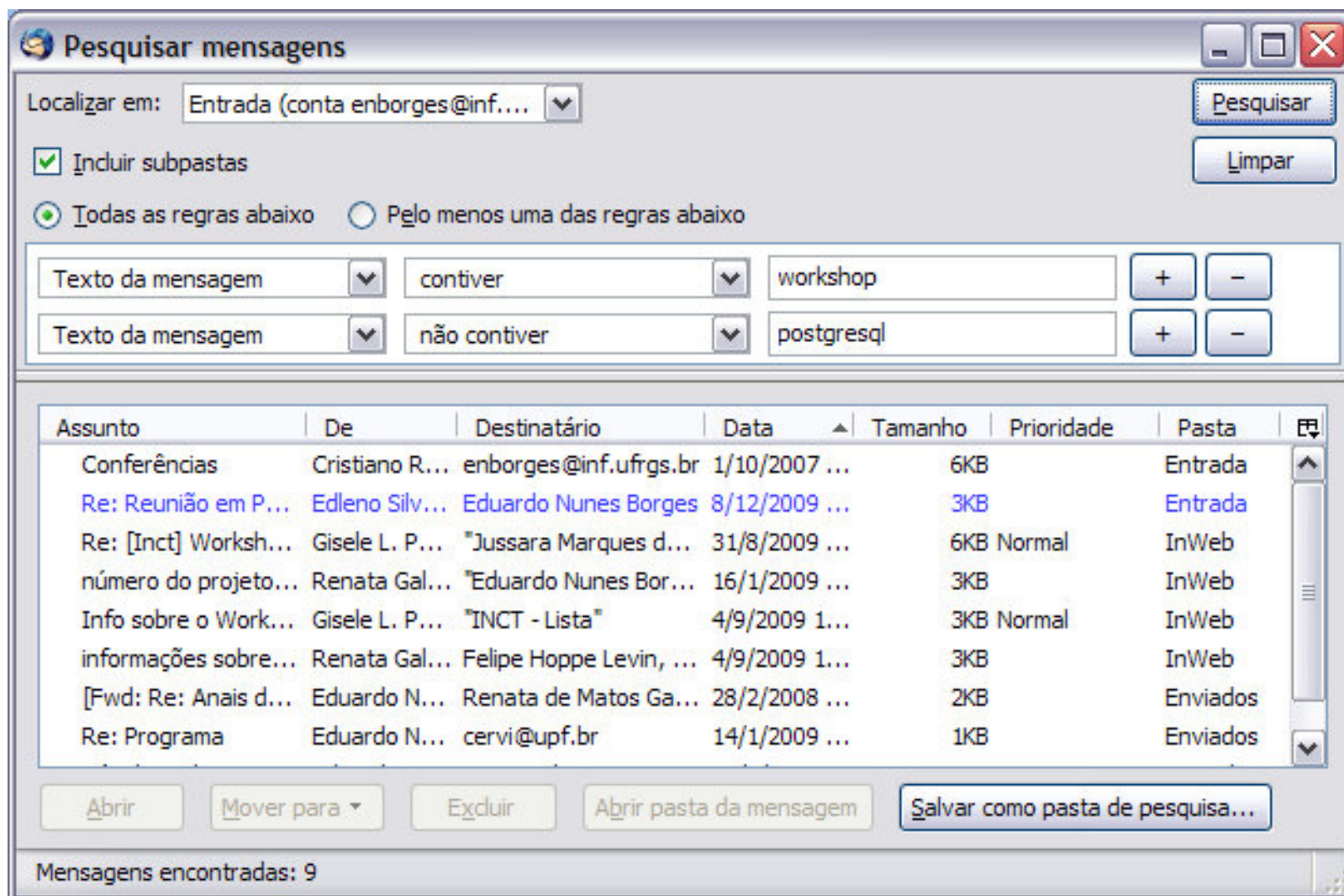
$$\{2,j\} \cup \{1,2,j\} = \{1,2,j\}$$

$$\{3,j\} \cap \{1,2,j\} = \{j\}$$

Modelo Booleano

- Formalismo claro, mas
 - Todos os documentos retornados possuem a mesma relevância
 - Não há *ranking*
 - Casamento exato
 - Dificuldade de expressar consultas utilizando operadores booleanos
 - Considerado o mais “fraco” dos modelos clássicos
- Extensão proposta: modelo booleano estendido [Salton, Fox & Wu, 1983]

Exemplo de Aplicação



Exercício

1. Considerando os seguintes documentos:

1. **xadrez.txt** = "O peão e o cavalo são peças de xadrez. O cavalo é o melhor do jogo."

2. **jogo.txt** = "A jogada envolveu a torre, o peão e o rei."

3. **rodeio.txt** = "O peão laçou o boi"

4. **fazenda.txt** = "Cavalo de rodeio!"

5. **policia.txt** = "Policiais o jogaram no xadrez."

removendo *stopwords* (lista: a, o, e, é, de, do, no, são) e considerando uma etapa de *stemming* utilizando o *Online Snowball stemmers* (<http://mazko.github.io/jssnowball/>) para língua portuguesa (language: *portuguese*), represente a coleção utilizando:

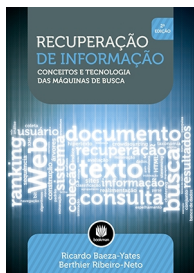
a) Matriz de incidências;

b) Índice baseado em arquivo invertido.

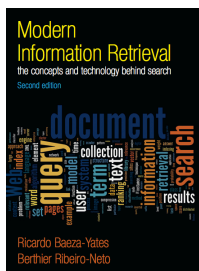
2. Realize a consulta booleana (**cavalo OR boi**) **AND NOT** **peão** para cada representação anterior.

3. Como pode ser otimizada a consulta (**peão OR cavalo OR torre**) **AND** (**jogo OR xadrez**)? Realize as operações passo a passo e indique as heurísticas utilizadas.

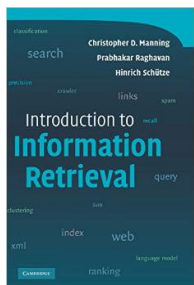
Referências



- Baeza-Yates, R.; Ribeiro-Neto, B. Recuperação de Informação: Conceitos e Tecnologia das Máquinas de Busca. 2 ed. Bookman, 2013.



- Baeza-Yates, R.; Ribeiro-Neto, B. Modern Information Retrieval. Wokingham, UK: Addison-Wesley, 2 ed., 2011.



- Manning, C. D.; Raghavan, P.; Schütze, H. Introduction to Information Retrieval. Cambridge University Press, 2008.

Online edition 2009: <http://nlp.stanford.edu/IR-book/>

Referências

- Salton, G.; Fox, E. A.; Wu, H. Extended Boolean information retrieval. Communications of the ACM, New York, v.26, n.11, p. 1022-1036, Nov. 1983.



Universidade Federal do Rio de Janeiro (UFRJ)
Departamento de Ciência da Computação (DCC)



Recuperação da Informação (MAB605)

Dúvidas?

Profa. Giseli Rabello Lopes
giseli@dcc.ufrj.br
CCMN - DCC - Sala E-2012

