

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



## Recuperação da Informação (MAB605)

Introdução

Profa. Giseli Rabello Lopes

### Roteiro

- Definições
- O processo de RI
- Componentes de um sistema de RI
- Considerações finais
- Referências

### Outras definições de RI

Associada a sistemas de recuperação de informação automática. Automática, em oposição ao manual e informação, em oposição aos dados ou fatos.

[Rijsbergen, 1979]

Trata da representação, armazenamento, organização e acesso a elementos de informação.

[Salton & McGill, 1983; Baeza-Yates & Ribeiro-Neto, 1999]

### Outras definições de RI

Dedicada a encontrar documentos relevantes, não simplesmente encontrar o casamento de padrões.

[Grossman & Frieder, 2004]

Encontrar material (geralmente documentos) de natureza não estruturada (geralmente texto) que satisfaça uma necessidade de informação nas grandes coleções (geralmente em servidores locais ou na internet).

[Manning et al., 2008]

## Introdução

- Inicialmente:
  - Informações textuais
  - Extensão: outros tipos de informação (sons, imagens, vídeos, etc.)



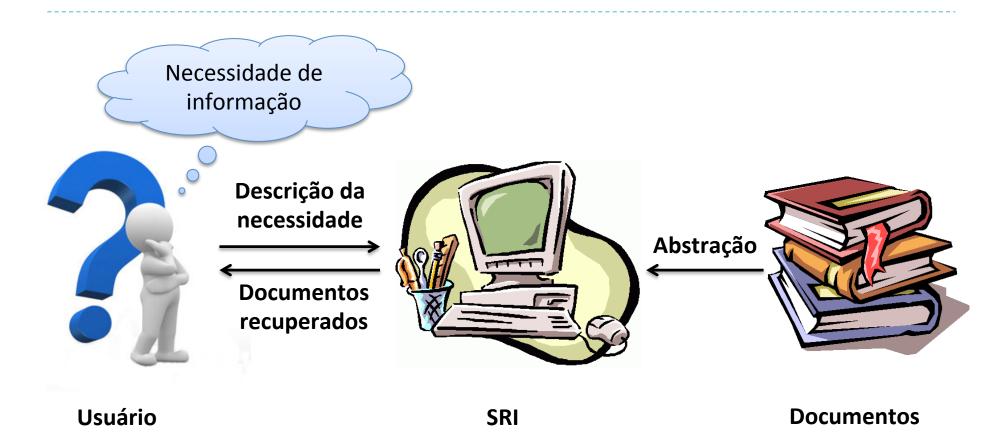
## Introdução

- RI costumava ser uma atividade na qual somente poucas pessoas estavam engajadas
- Atualmente, centenas de milhões de pessoas realizam atividades de RI diariamente
  - Ex.: Utilizar um motor de busca na Web
- RI está se tornando rapidamente a forma dominante de acesso à informação
  - Ultrapassando o estilo de busca em bancos de dados tradicionais

# Introdução

Recuperação de Dados	Recuperação de Informações
Dados estruturados	Dados não estruturados
Consulta complexa e Estruturada (linguagem de consulta)	Consulta através de palavras- chave
Casamento exato	Casamento parcial ou melhor casamento
Retornados ou não retornados	Ranking dos resultados

## O processo de RI



Adaptado de: [Wives, 2001]

### 1. O usuário tem uma necessidade de informação

Quais doenças ou problemas médicos são suspeitos de estarem relacionados ao fumo ativo ou passivo?



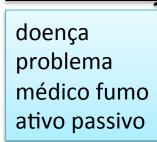
Usuário

# 2. O usuário tipicamente precisa traduzir sua necessidade de informações em forma de uma consulta (*keywords*)

Quais doenças ou problemas médicos são suspeitos de estarem relacionados ao fumo ativo ou passivo?

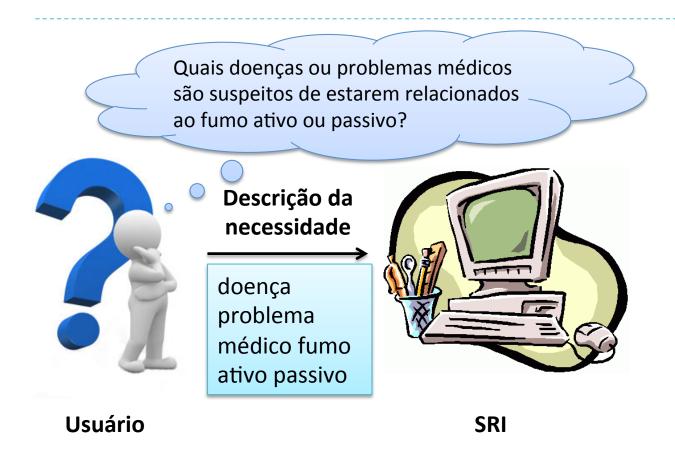
Descrição da

necessidade



Usuário

### 3. A consulta é submetida a um sistema de IR



# 4. O sistema de IR processa esta consulta e devolve uma lista de documentos classificada em ordem decrescente de relevância

Quais doenças ou problemas médicos são suspeitos de estarem relacionados ao fumo ativo ou passivo?



Descrição da necessidade

## **Documentos** recuperados

Fumantes passivos - Hospital Israelita Albert Einstein

www.einstein.br/einstein-saude/vida-saudavel/.../fumantes-passivos.aspx ▼ Boletim médico · Glossário de termos médicos ... O fumo passivo é a maior causa das doenças respiratórias nas crianças ... Esses bebês também têm mais chances de desenvolver problemas respiratórios crônicos, ... O câncer de pulmão é a mais perigosa doença respiratória que acomete fumantes ativos e passivos.



### Doenças causadas pelo cigarro - tabagismo, fumante ativo ...

www.suapesquisa.com/ecologiasaude/doencas\_cigarro.htm ~ Principais doenças causadas pelo cigarro, fumante ativo e passivo, tabagismo, ... Os médicos afirmam que, o corpo humano leva de um a dois anos para limpar ...

### Os perigos de ser um fumante passivo - Boa Saúde

www.boasaude.com.br/.../os-perigos-de-ser-um-fumante-passivo.html ▼
Então veja aqui os lançamentos da Biblioteca médica. ... O tabagismo passivo
corresponde à exposição de pessoas não fumantes ao ar contaminado ... o tabagismo
passivo pode causar as mesmas doenças provocadas pelo tabagismo ativo, ... nas
crianças e nas pessoas com problemas respiratórios e cardiovasculares.

### O fumo interfere na circulação sanguínea, causando ...

clinicaviarengo.com/.../o-fumo-interfere-na-circulacao-sanguinea-causan... ▼
29 de out de 2012 - Para você ter uma ideia, hoje, os portadores de doenças
vasculares causadas pelo ... De acordo com o médico, fumantes ativos e passivos
apresentam um risco ... Em relação aos problemas vasculares, essa fumaça produz.



Abstração



SRI



### Sistemas de RI

- Definição formal [Baeza-Yates & Ribeiro-Neto, 1999]
  - $-SRI = [D, Q, M, R(q_i, d_i)]$
  - -D é um conjunto de documentos (coleção)
  - Q é um conjunto de representações das necessidades de informação do usuário (consulta)
  - -M é um modelo
    - Representação dos documentos e consultas
    - Relacionamentos entre eles
  - -R é uma função de *ranking* que associa um valor passível de ordenação a um par composto por uma consulta  $q_i$  e um documento  $d_i$

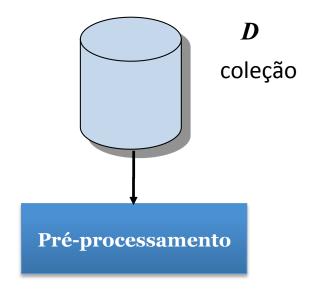
Tokenização

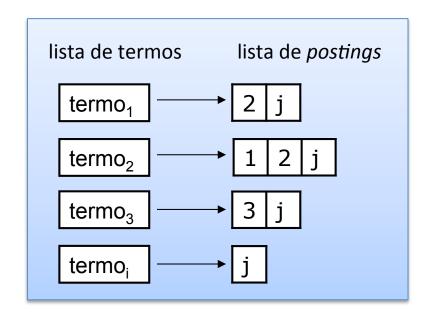
Normalização

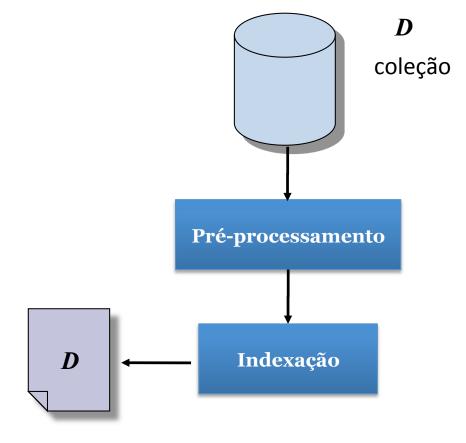
Remoção de palavras irrelevantes (stopwords)

Remoção de afixos (stemming)

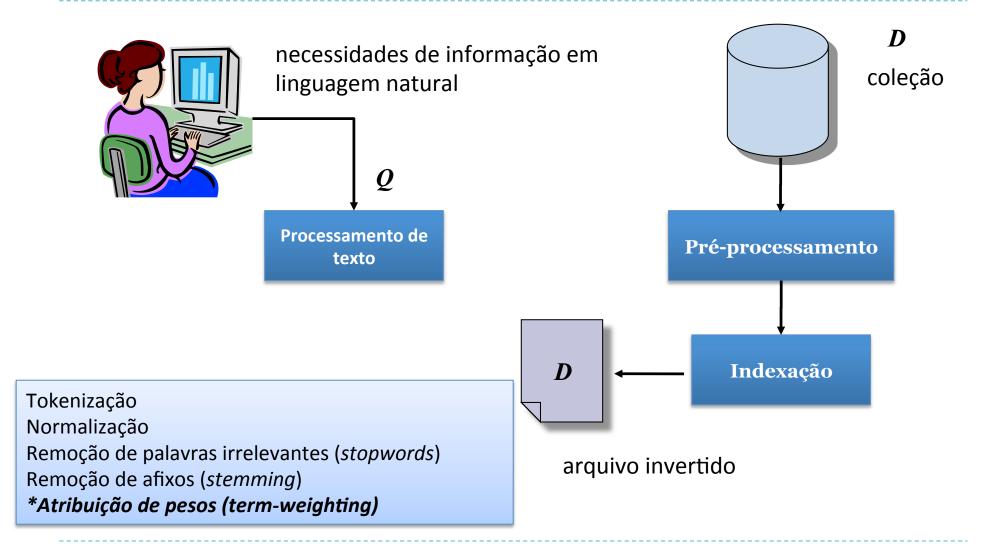
Atribuição de pesos (term-weighting)

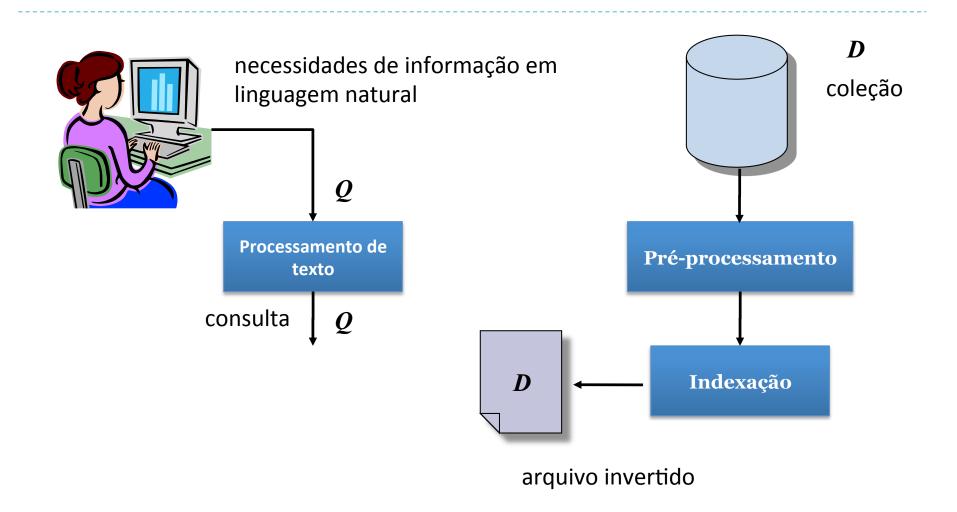


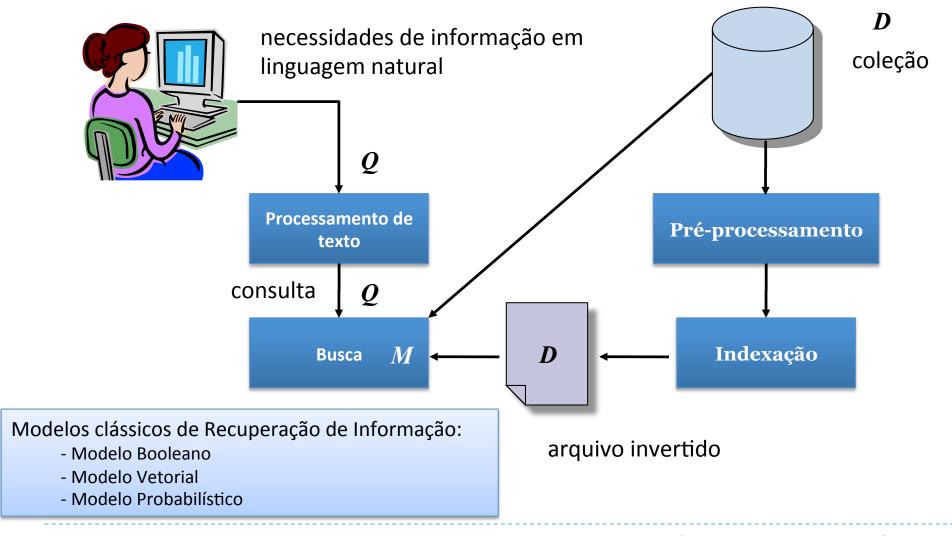


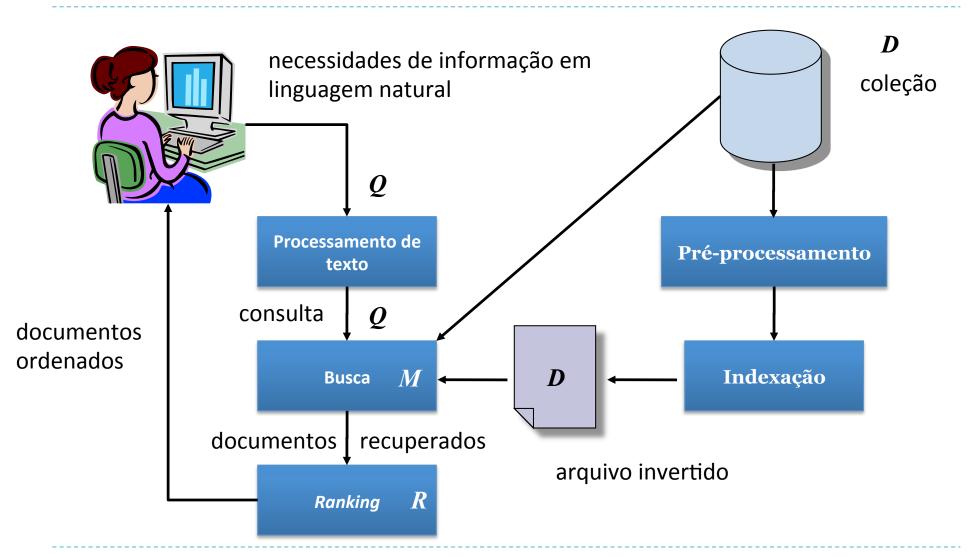


arquivo invertido









### Considerações finais

[Baeza-Yates & Ribeiro-Neto, 2013]

- RI trata da representação, armazenamento, organização e acesso de itens de informação
  - Tipos de itens de informação: documentos, páginas Web, catálogos online, registros estruturados e semiestruturados, objetos multimídia
- Objetivos iniciais da área de RI:
  - Indexação de textos e busca por documentos úteis em uma coleção

### Considerações finais

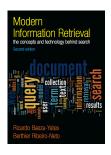
[Baeza-Yates & Ribeiro-Neto, 2013]

- Atualmente, pesquisas em RI incluem:
  - Modelagem, busca na Web, classificação de texto, buscas semânticas, arquitetura de sistemas, interfaces de usuário, visualização de dados, filtragem e linguagens

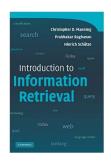
### Referências



 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: <a href="http://nlp.stanford.edu/IR-book/">http://nlp.stanford.edu/IR-book/</a>

### Referências

- Grossman, D. A.; Frieder, O. Information retrieval: algorithms and heuristics. 2nd ed. Dordrecht: Springer, c2004. 332p.
- Rijsbergen, C. J. Information retrieval. London: Butterworths, 1979.
- Salton, G.; Mcgill, M. J. Introduction to Modern Information Retrieval. New York: McGraw-Hill Book. 1983. 448 p.
- 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.
- Wives, L. K. Técnicas de descoberta de conhecimento em textos aplicadas à inteligência competitiva. 2001. Exame de Qualificação (doutorado em Ciência da Computação) -- Instituto de Informática, UFRGS, Porto Alegre.



### 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

