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

DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO - UFRJ -

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.)

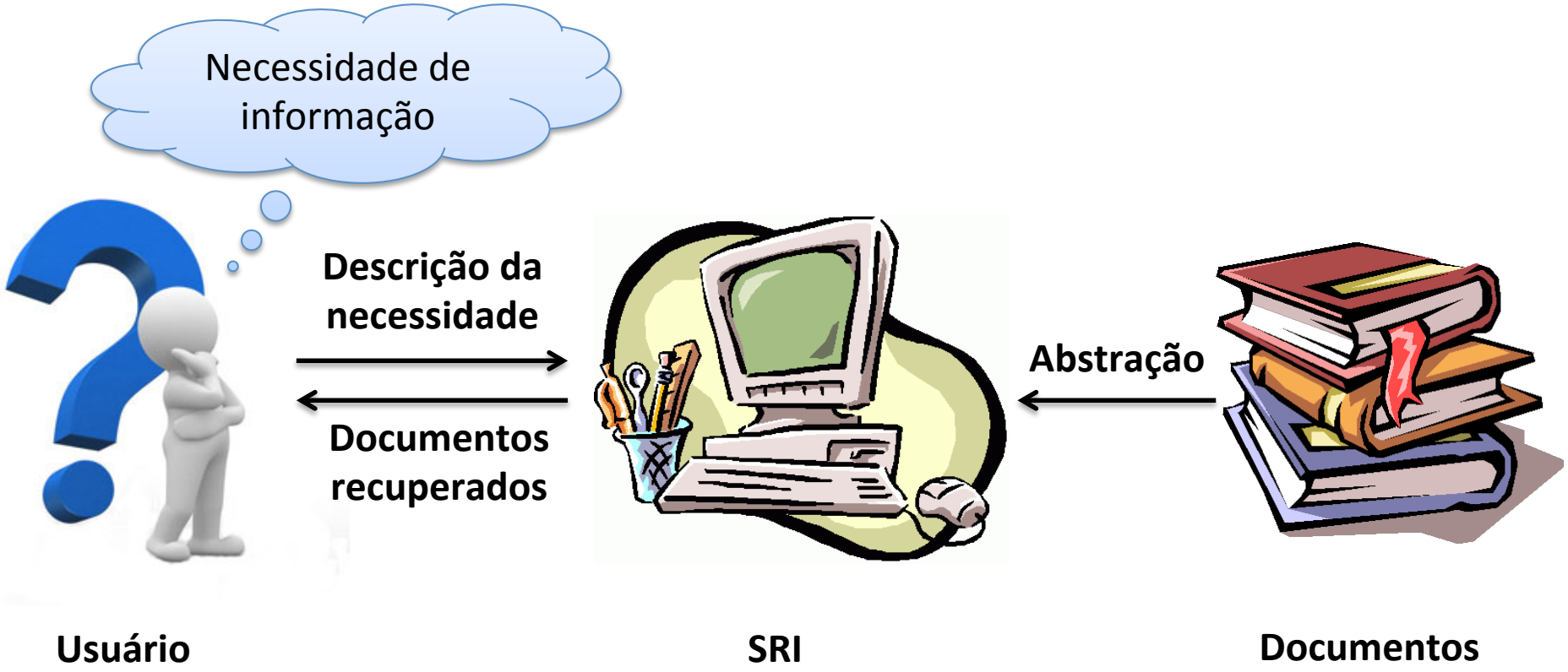
Introdução

- 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

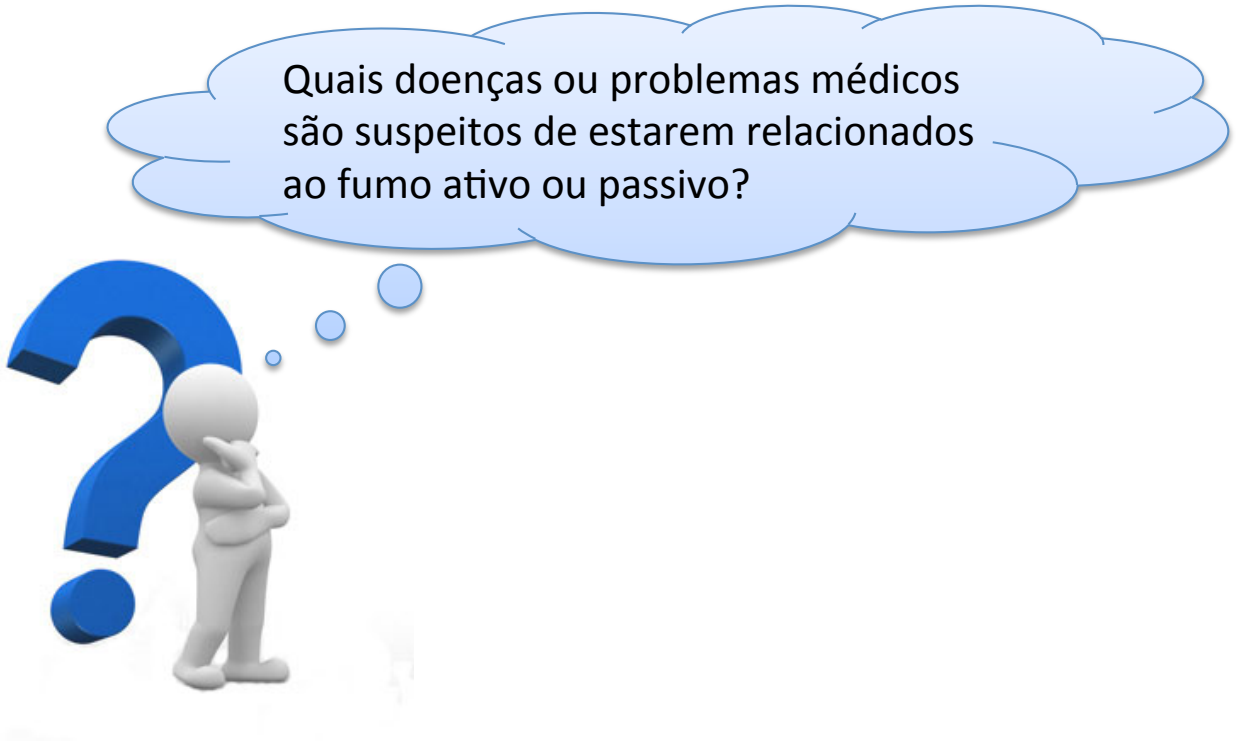
Introdução

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

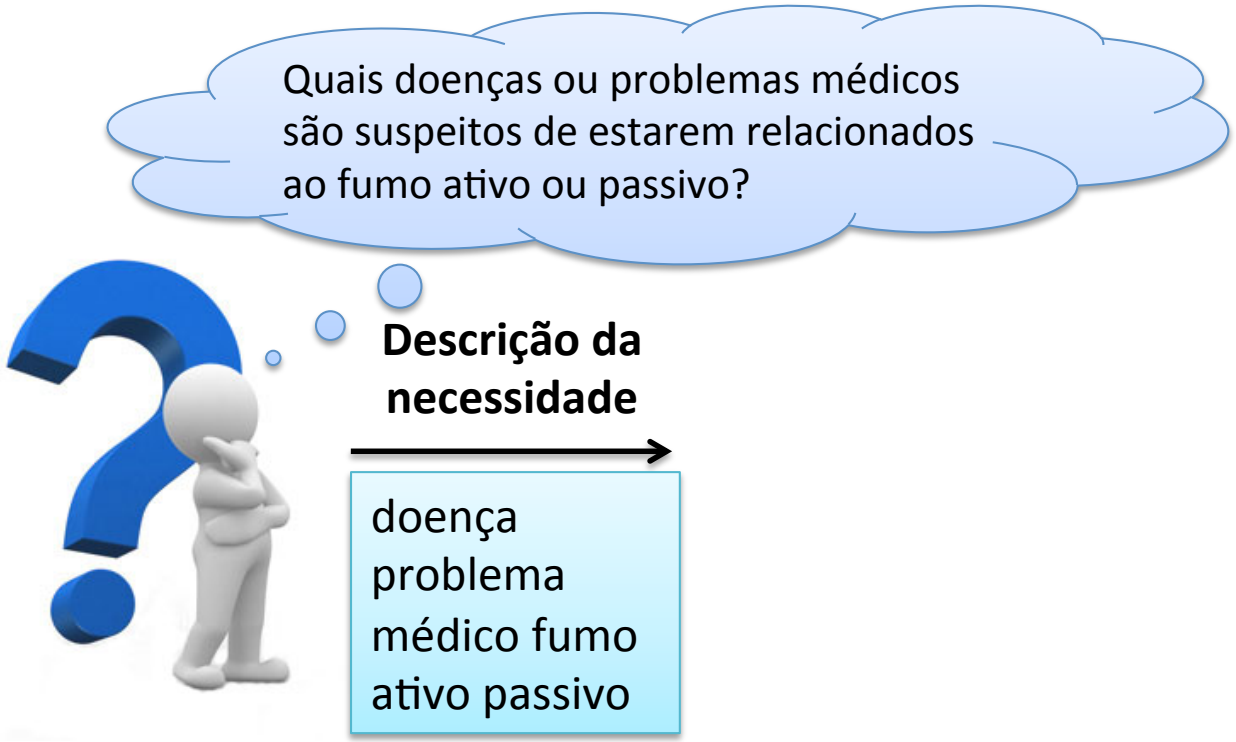


1. O usuário tem uma necessidade de informação

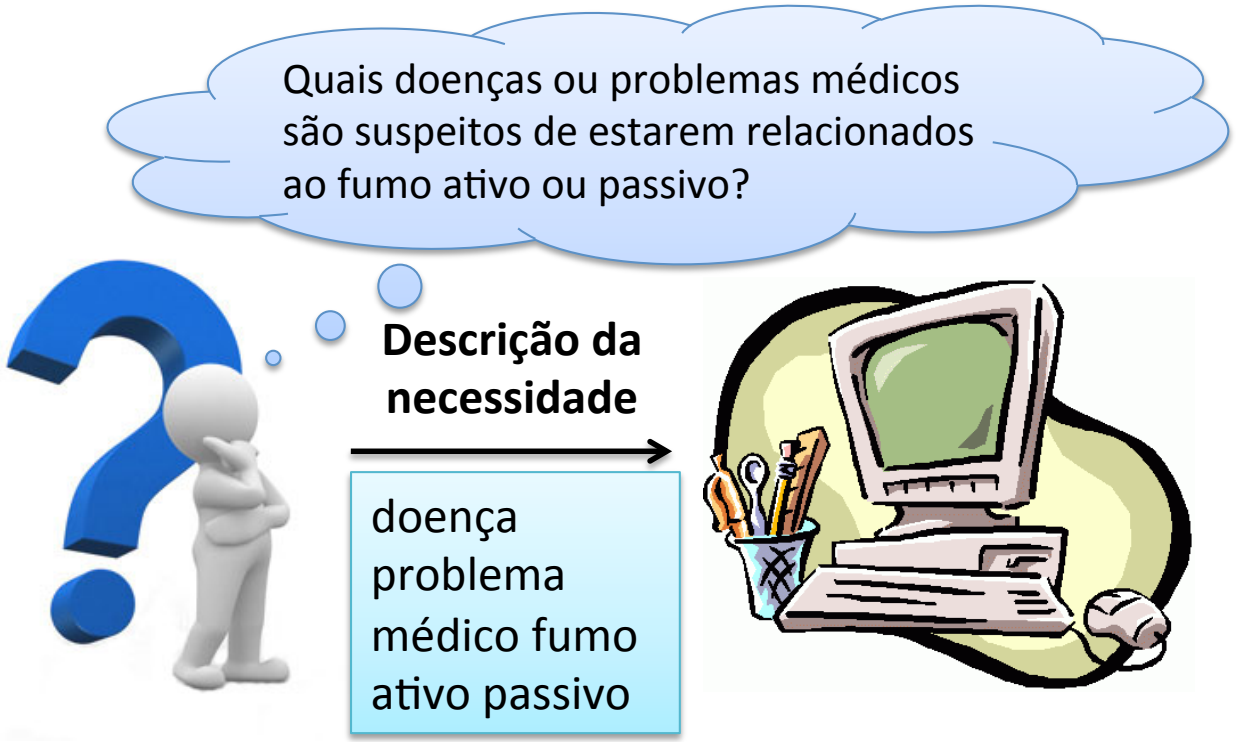


Usuário

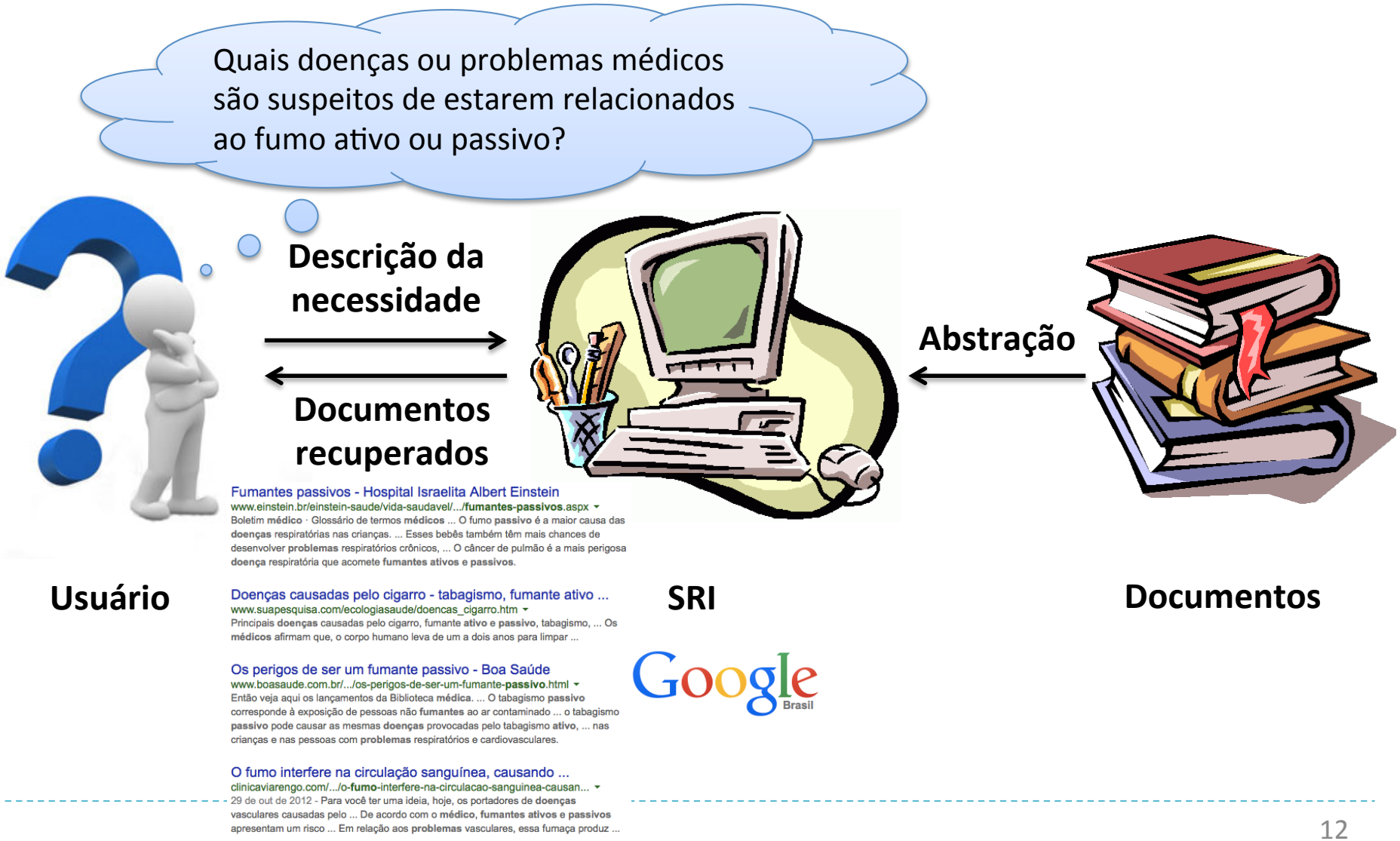
2. O usuário tipicamente precisa traduzir sua necessidade de informações em forma de uma consulta (*keywords*)



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

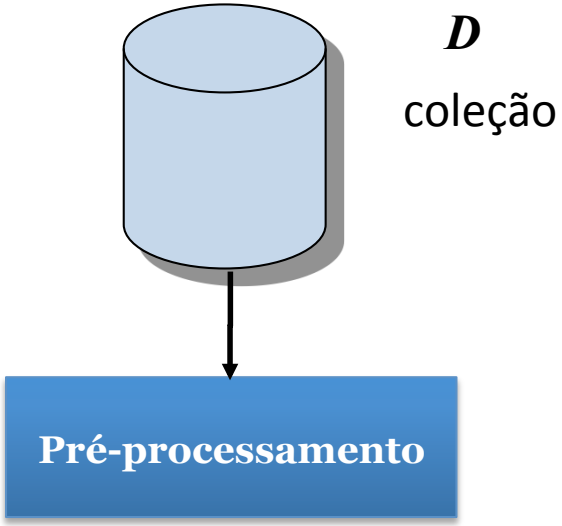


Sistemas de RI

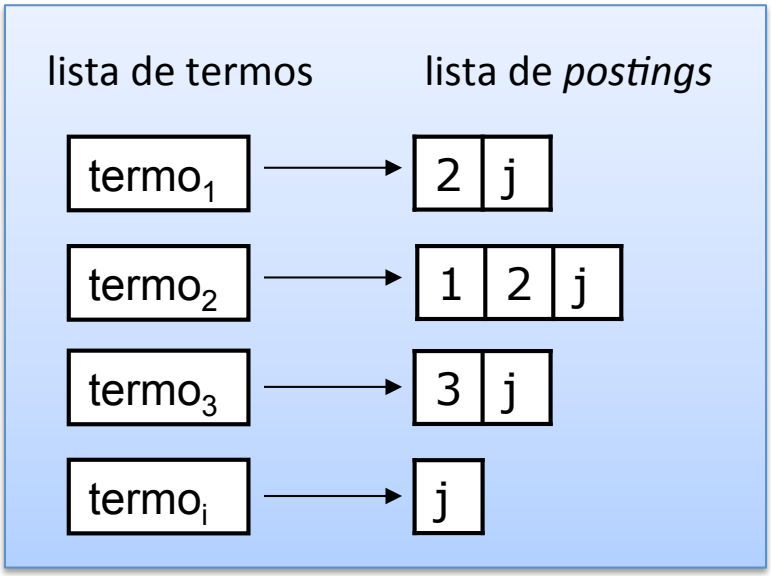
- Definição formal [Baeza-Yates & Ribeiro-Neto, 1999]
  - $SRI = [D, Q, M, R(q_i, d_j)]$
  - D é um conjunto de documentos (coleção)
  - Q é um conjunto de representações das necessidades de informação do usuário (consulta)
  - M é um modelo
    - Representação dos documentos e consultas
    - Relacionamentos 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_j

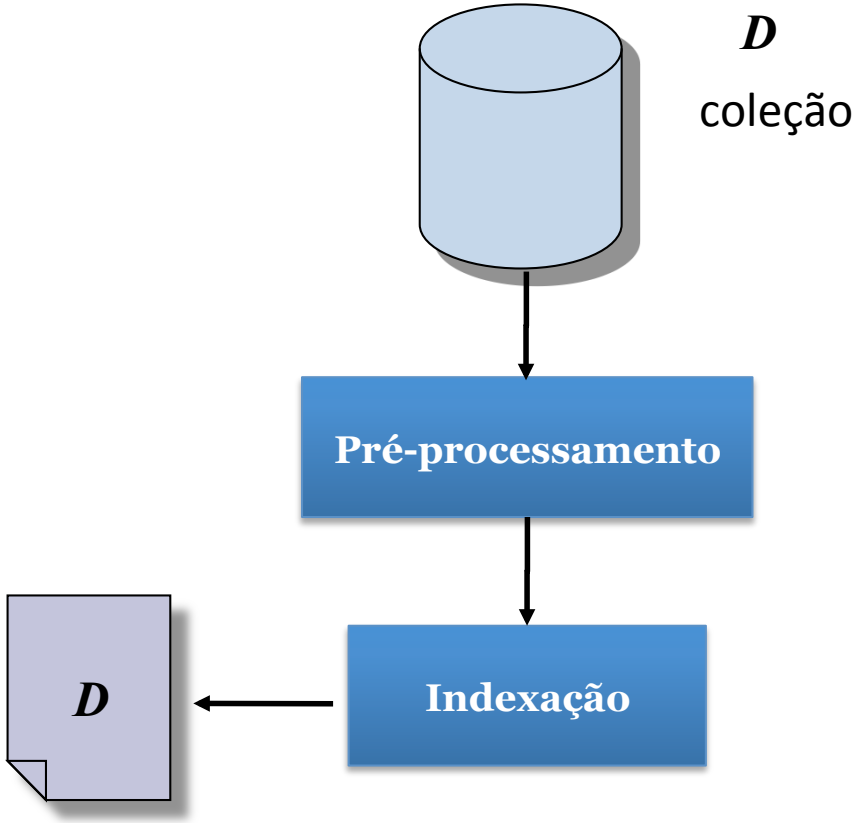
Componentes de um Sistemas de RI

Tokenização
Normalização
Remoção de palavras irrelevantes (*stopwords*)
Remoção de afixos (*stemming*)
Atribuição de pesos (*term-weighting*)



Componentes de um Sistemas de RI





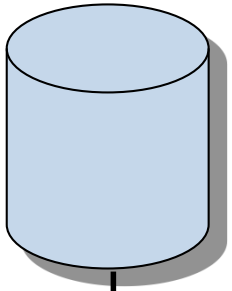
arquivo invertido

Componentes de um Sistemas de RI

necessidades de informação em linguagem natural

Q

Processamento de texto

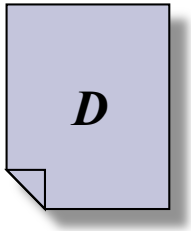


D

coleção

Pré-processamento

Indexação

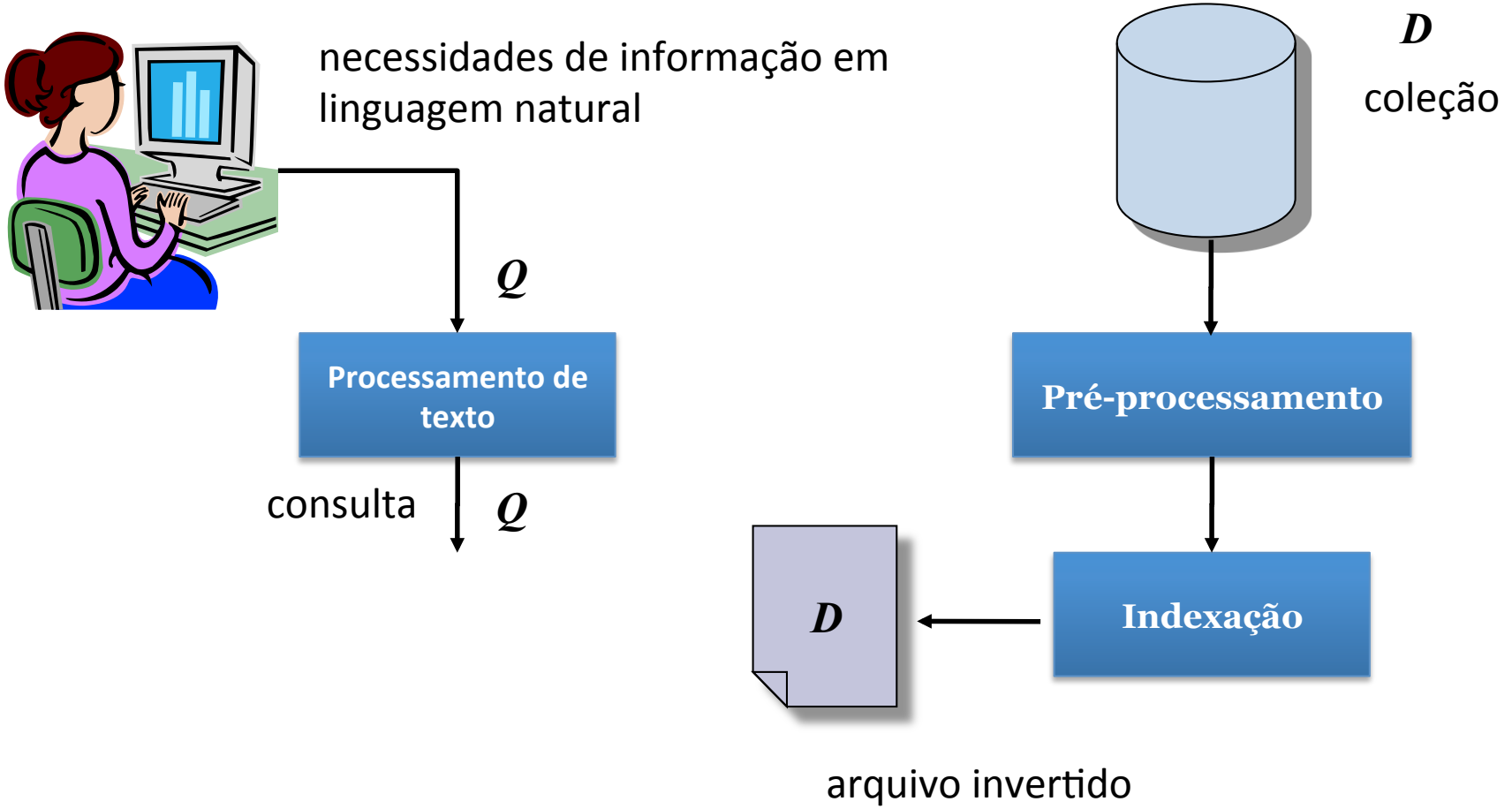


D

arquivo invertido

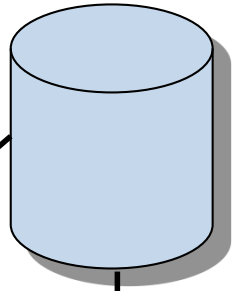
Tokenização
Normalização
Remoção de palavras irrelevantes (*stopwords*)
Remoção de afixos (*stemming*)
Atribuição de pesos (term-weighting)

Componentes de um Sistemas de RI



Componentes de um Sistemas de RI

necessidades de informação em linguagem natural



D

coleção

Q

Processamento de texto

Pré-processamento

consulta Q

Busca M

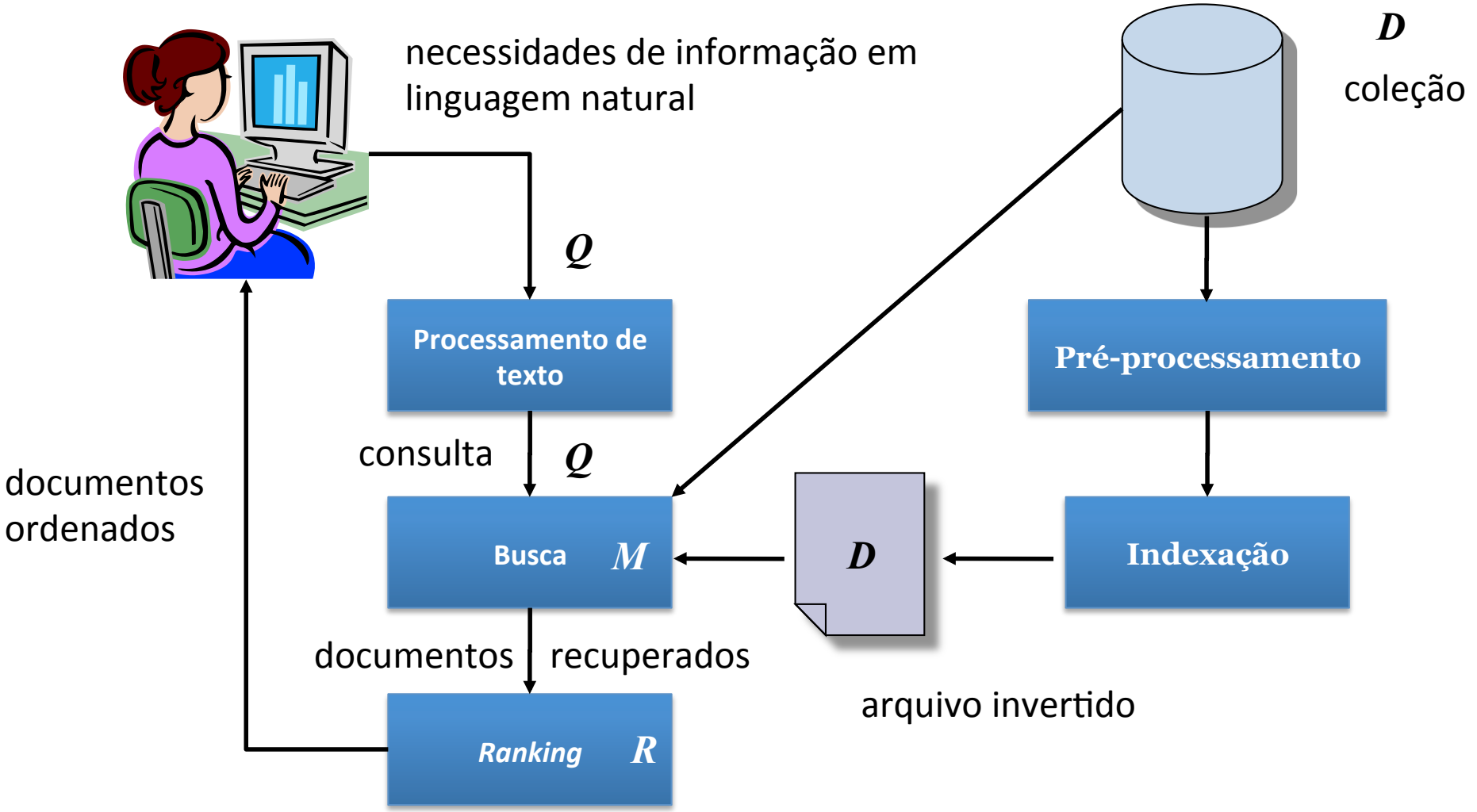
D

Indexação

arquivo invertido

Modelos clássicos de Recuperação de Informação:
- Modelo Booleano
- Modelo Vetorial
- Modelo Probabilístico

Componentes de um Sistemas de RI



Considerações finais

[Baeza-Yates & Ribeiro-Neto, 2013]

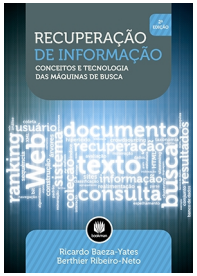
- RI trata da representação, armazenamento, organização e acesso de itens de informação
  - Tipos de itens de informação: documentos, páginas Web, catálogos online, registros estruturados e semiestruturados, objetos multimídia
- Objetivos iniciais da área de RI:
  - Indexação de textos e busca por documentos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

Considerações finais

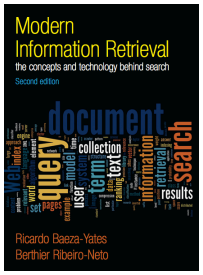
[Baeza-Yates & Ribeiro-Neto, 2013]

- 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

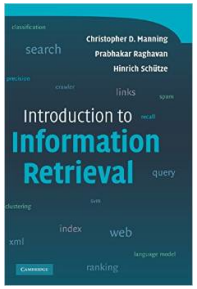
Referências



- 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: http://nlp.stanford.edu/IR-book/

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)

DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO - UFRJ -

Recuperação da Informação (MAB605)

Dúvidas?

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