Indexação e Modelos Clássicos

Wendel Melo
Faculdade de Computação
Universidade Federal de Uberlândia

Recuperação da Informação

Adaptado do Material da Profª Vanessa Braganholo - IC/UFF

Construção de um Sistema de Busca

- **1ª Etapa: Definir a base de documentos:**

Construção de um Sistema de Busca

- **1ª Etapa: Definir a base de documentos:**
  - Na maioria dos casos, a base é textual;
  - Em muitos casos, pode ser necessário construir algum tipo de aplicação para coletar os documentos em repositórios ou na internet;
  - Define-se o modelo de texto e quais elementos poderão ser recuperados: Palavras? Parágrafos? Páginas? Documento inteiro?
  - Define-se as operações sobre a base, como remoção de palavras vazias (*stopwords*), extração de radicais, etc.

Construção de um Sistema de Busca

- **1ª Etapa: Definir a base de documentos:**
- **2ª Etapa: Indexação da base de documentos**
  - O Objetivo é permitir busca rápida sobre a base;
  - Geralmente são utilizados **índices invertidos**;

Índice Invertido

- **Objetivo**: construir uma estrutura de dados que permita rapidamente saber quais documentos contém um determinado termo.

Índice Invertido

- **Objetivo**: construir uma estrutura de dados que permita rapidamente saber quais documentos contém um determinado termo.

- Lista-se todas as palavras distintas da base de dados (após operações como remoção de *stopwords* (palavras sem significado próprio relevante) e extração de radicais);

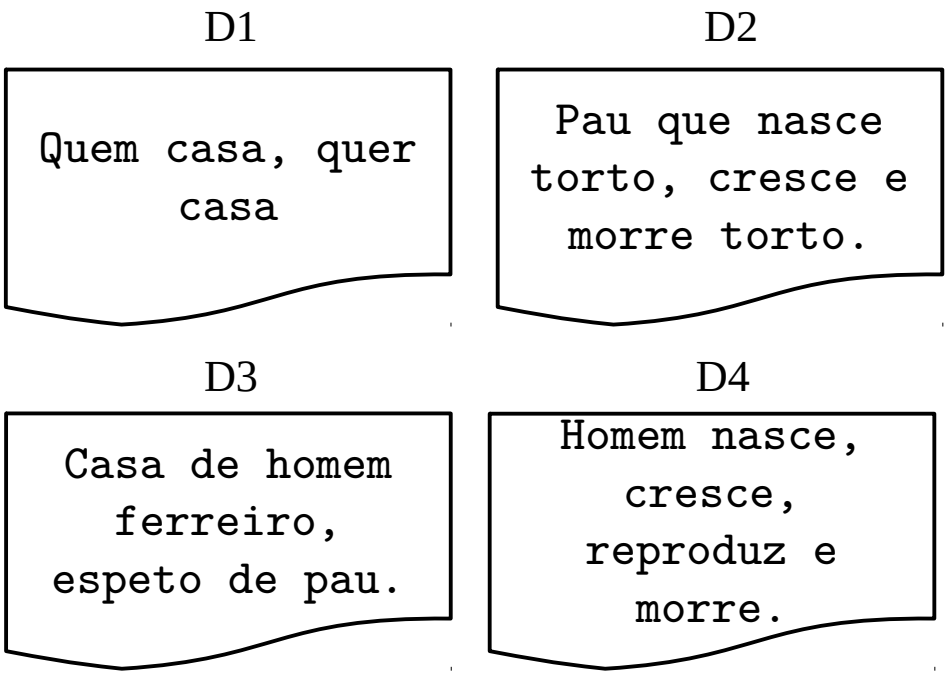
Índice Invertido

- **Objetivo**: construir uma estrutura de dados que permita rapidamente saber quais documentos contém um determinado termo.

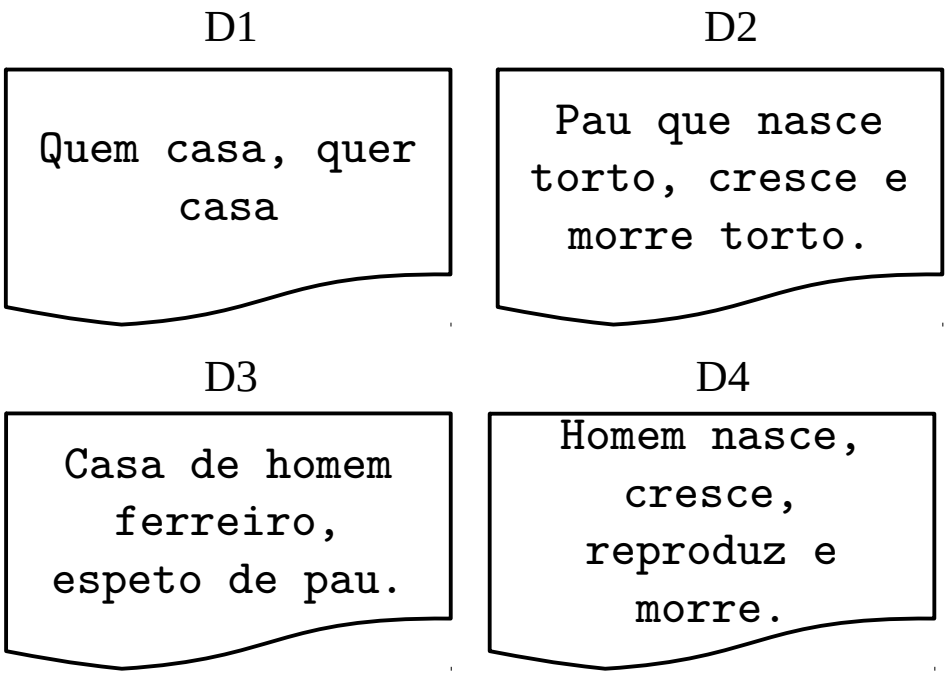
- Lista-se todas as palavras distintas da base de dados (após operações como remoção de *stopwords* (palavras sem significado próprio relevante) e extração de radicais);

- Para cada palavra da base, listamos os documentos em que a mesma aparece. Para cada documento D onde a palavra em questão aparece, registramos também a quantidade de vezes em que a palavra aparece em D

Índice Invertido - Exemplo



Índice Invertido - Exemplo



Temos 4 documentos: D1, D2, D3 e D4

Desconsideramos as *stopwords*: de, e, que, quem

Para simplificar, não extrai-remos os radicais.

Índice Invertido - Exemplo

D1

Quem casa, quer casa

D2

Pau que nasce torto, cresce e morre torto.

D3

Casa de homem ferreiro, espeto de pau.

D4

Homem nasce, cresce, reproduz e morre.

casa	→	(D1,2) (D3,1)
cresce	→	(D2,1) (D4,1)
espeto	→	(D3,1)
ferreiro	→	(D3,1)
homem	→	(D3,1) (D4,1)
morre	→	(D2,1) (D4,1)
nasce	→	(D2,1) (D4,1)
pau	→	(D2,1) (D3,1)
quer	→	(D1,1)
reproduz	→	(D4,1)
torto	→	(D2,2)

Índice Invertido - Exemplo

D1

Quem casa, quer casa

D2

Pau que nasce torto, cresce e morre torto.

D3

Casa de homem ferreiro, espeto de pau.

D4

Homem nasce, cresce, reproduz e morre.

O termo “casa” aparece no documento D1 duas vezes, e no documento D3 uma vez.

casa	→	(D1,2) (D3,1)
cresce	→	(D2,1) (D4,1)
espeto	→	(D3,1)
ferreiro	→	(D3,1)
homem	→	(D3,1) (D4,1)
morre	→	(D2,1) (D4,1)
nasce	→	(D2,1) (D4,1)
pau	→	(D2,1) (D3,1)
quer	→	(D1,1)
reproduz	→	(D4,1)
torto	→	(D2,2)

Índice Invertido

- Adicionalmente, o índice invertido pode também armazenar a posição em que os termos ocorrem nos arquivos.

Índice Invertido

- Adicionalmente, o índice invertido pode também armazenar a posição em que os termos ocorrem nos arquivos.

- Localizar rapidamente um determinado termo no índice invertido é de suma importância, por essa razão, estruturas que permitem acesso rápido como árvores ou tabelas *hash* (dicionários /mapas) costumam ser utilizadas.

Índice Invertido

- Adicionalmente, o índice invertido pode também armazenar a posição em que os termos ocorrem nos arquivos.

- Localizar rapidamente um determinado termo no índice invertido é de suma importância, por essa razão, estruturas que permitem acesso rápido como árvores ou tabelas *hash* (dicionários /mapas) costumam ser utilizadas.

- O conjunto de termos indexados é denominado **vocabulário da base**.

Construção de um Sistema de Busca

- **1ª Etapa: Definir a base de documentos**
- **2ª Etapa: Indexação da base de documentos**
- **3ª Etapa: Definição de um modelo para elaboração de consultas e para responder as mesmas**
  - O usuário poderá especificar uma consulta que será sintaticamente analisada e expandida com sinônimos e/ou termos relacionados;
  - A partir do índice, obtém-se uma lista com um conjunto de documentos a serem recuperados;
  - Os documentos são **ranqueados** segundo sua suposta relevância ao usuário e apresentados seguindo a ordem do ranking.

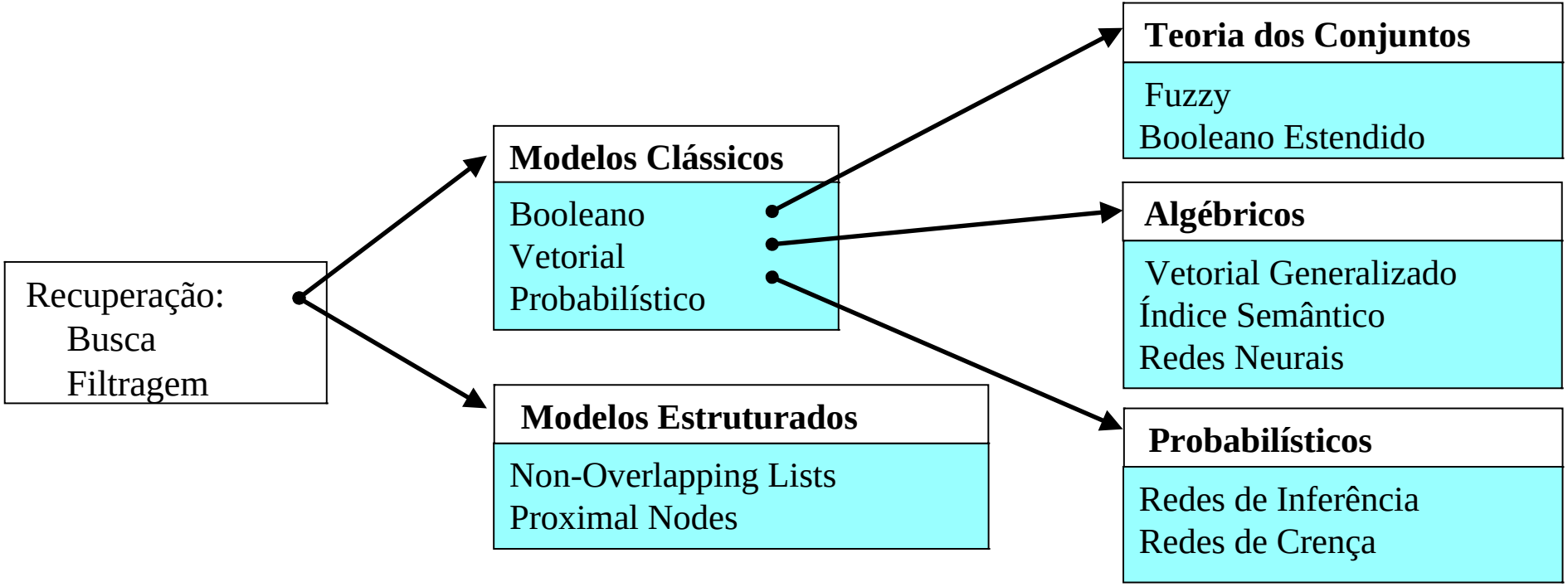
Modelos de RI

- Parte principal dos sistemas de RI;
- Determina como documentos, termos e consultas serão tratados afim de determinar relevância entre documentos e consultas e produzir ranqueamento;

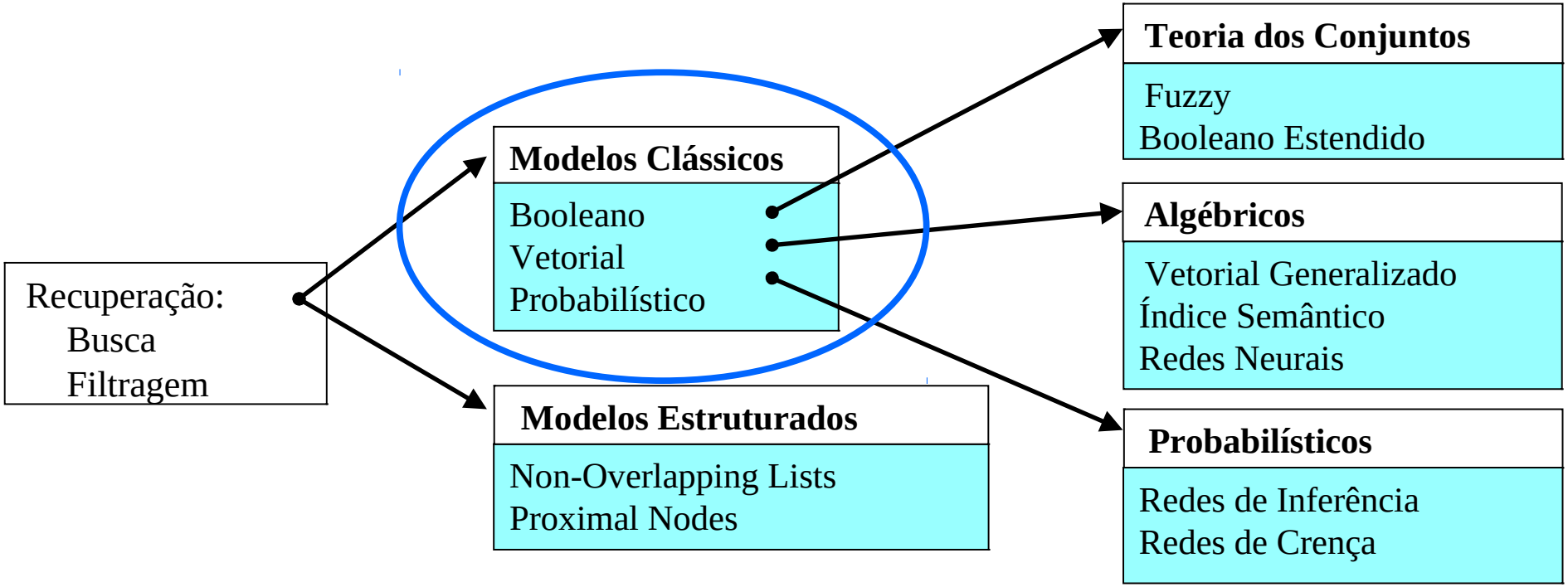
Modelos de RI

- A definição formal de um modelo de RI envolve quatro elementos principais (D, Q, F, $R(q_i, d_j)$), onde:
- D: conjunto das representações dos documentos da base;
- Q: conjunto das representações das consultas;
- F: *framework* para a modelagem dos documentos, consultas e o relacionamento entre ambos;
- $R(q_i, d_j)$: função de ordenamento que associa a cada consulta $q_i \in Q$ e a cada documento $d_j \in D$, um **grau de similaridade** entre o documento e a consulta.

Modelos de RI



Modelos de RI



Modelos Clássicos de RI

- Foco da disciplina;
- São destinados a RI não estruturada;
- Três modelos fundamentais:
  - Modelo Booleano
  - Modelo Vetorial
  - Modelo Probabilístico
- Variações dos 3 modelos clássicos surgiram adotando técnicas mais sofisticadas

Modelos Clássicos de RI

- Os modelos partem da ideia de atribuição de pesos aos termos que compõem os documentos;

Modelos Clássicos de RI

- Os modelos partem da ideia de atribuição de pesos aos termos que compõem os documentos;

- Os modelos booleano e probabilístico adotam pesos binários para indicar presença ou ausência de um termo;

Modelos Clássicos de RI

- Os modelos partem da ideia de atribuição de pesos aos termos que compõem os documentos;
- Os modelos booleano e probabilístico adotam pesos binários para indicar presença ou ausência de um termo;
- O modelo vetorial pode adotar esquemas de ponderação mais sofisticados, onde o peso de um termo em um documento pode ser proporcional à sua importância para descrever o documento e a raridade na base.