INSTITUTO FEDERAL
MINAS GERAIS

RECUPERAÇÃO DE INFORMAÇÃO

Profa. Patrícia Proença

patricia.proenca@ifmg.edu.br

INSTITUTO FEDERAL
MINAS GERAIS

ATENÇÃO!!!

- O material a seguir é uma videoaula apresentada pela professora PATRÍCIA APARECIDA PROENÇA AVILA, como material pedagógico do IFMG, dentro de suas atividades curriculares ofertadas em ambiente virtual de aprendizagem. Seu uso, cópia e ou divulgação em parte ou no todo, por quaisquer meios existentes ou que vierem a ser desenvolvidos, somente poderá ser feito, mediante autorização expressa deste docente e do IFMG. Caso contrário, estarão sujeitos às penalidades legais vigentes".

- Conforme Art. 2º§1º da Nota Técnica nº 1/2020/PROEN/Reitoria/IFMG (SEI 0605498, Processo nº 23208.002340/2020-04

Modelos de RI

Tópicos

- Modelagem em RI;
- Caracterização de um modelo de RI;
- Taxonomia de modelos de RI;
- Recuperação de informação clássica;

Breve resumo da aula anterior

Aula anterior

- O objetivo da área de estudo conhecida como Recuperação de Informação é:
  - Prover aos usuários o **acesso fácil** às informações de seu interesse
- A diferença entre os objetivos iniciais da área e como esses objetivos mudaram com o advento da Web;
- Breve histórico.

Aula anterior

- O **problema de RI** está ligado basicamente a:
  - Como **extrair** as informações dos documentos?
  - Como utilizar tais informações para decidir sobre a sua relevância?
- Sistema de RI pode ser dividido em **seis módulos**:
  - 1) **Obtenção e coleção** de documentos;
  - 2) **Indexação** dos documentos;
  - 3) **Consulta** do usuário;
  - 4) **Recuperação** de documentos;
  - 5) **Ranqueamento** dos documentos;
  - 6) **Apresentação** para o usuário.

Modelagem em RI

Modelagem em RI

- Um sistema de recuperação de informação pode ser visto como
  - a parte do sistema de informação responsável pelo **armazenamento ordenado dos documentos em base de dados**,
  - e sua posterior **recuperação** para responder a **consulta do usuário**.
- Todo SRI adota um **modelo computacional de recuperação de informação** que determina o modo de operação do sistema.

Modelagem em RI

- Modelagem em RI é um processo complexo que tem o **objetivo de produzir uma função de ranqueamento**, ou seja, uma função que atribui **valores (pesos) a documentos em relação a uma consulta**;

- Esse processo pode ser dividido em duas tarefas principais:
  - 1) Concepção de um **sistema lógico para representar documentos e consultas**;
  - 2) A **definição de uma função de ranqueamento** que computa o grau de similaridade de cada documento em relação à consulta dada.

Caracterização de um modelo de RI

Modelagem e Ranqueamento

- Sistemas de RI geralmente adotam **termos de indexação** para **indexar e recuperar** documentos.
  - recuperar respostas para uma consulta qualquer tem que lidar com um problema central - **prever** quais **documentos os** usuários irão considerar **relevantes** e quais serão **irrelevantes**.
- Existe um grau de **incerteza** ou de **imprecisão** envolvido nesse processo, pois **dois usuários podem discordar** sobre o que é e o que não é relevante;

Modelagem e Ranqueamento

- Visando amenizar esse problema o sistema de RI **implementa um algoritmo preditivo que almeja aproximar-se da opinião de uma grande fração dos usuários** quanto a relevância dos resultados de uma grande fração de consultas:
  - Algoritmo é a **função de ranqueamento**.

Caracterização de um modelo de RI

- As premissas fundamentais que formam a base de um algoritmo de ranqueamento determinam o modelo de RI.

- A Caracterização de um modelo de RI é a seguinte:

  - Um modelo de RI é uma quádrupla $[D,Q,F,R(q_i,d_j)]$

Caracterização de um modelo de RI

- Um modelo de RI é uma quádrupla $[D,Q,F,R(q_i,d_j)]$ onde:

- 1. D é um **conjunto** composto por visões lógicas (ou representações) **dos documentos da coleção**.

- 2. Q é um **conjunto** composto por visões lógicas (ou representações) **das necessidades de informação dos usuários**. Essas representações são chamadas de **consultas**.

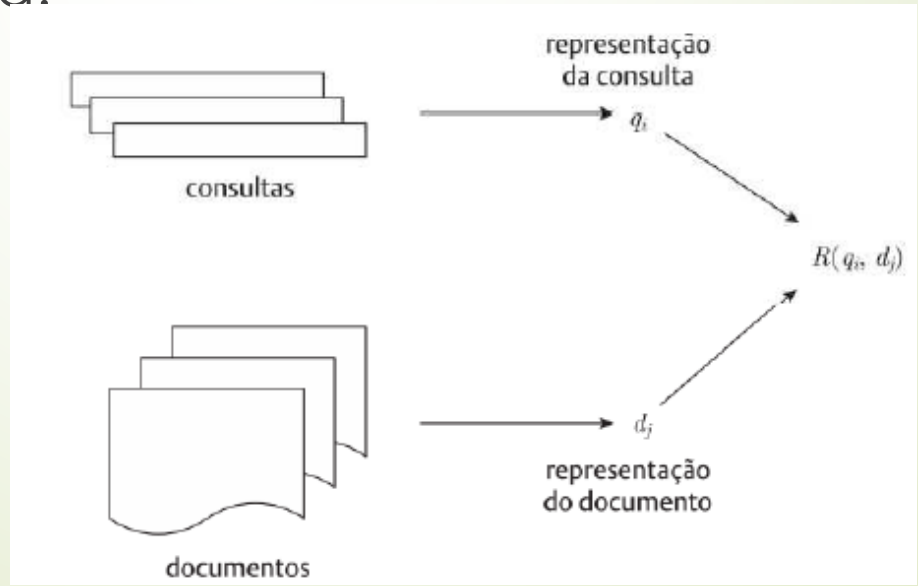
Caracterização de um modelo de RI

- 3. F é um sistema lógico usado para **modelar as representações dos documentos, das consultas e de seus relacionamentos**, como conjuntos e relações Booleanas;

- 4. $R(q_i,d_j)$ é uma **função de ranqueamento** que associa um número real à representação de uma consulta $q_i \in Q$ e à representação de um documento $d_j \in D$. Esse ranking define um ordenamento entre os documentos em relação à consulta q_i.

Caracterização de um modelo de RI – Função de ranqueamento

- Dadas as representações da consulta e dos documentos, como q_i e d_j, a função de ranqueamento $R(q_i, d_j)$ atribui um grau (número real) ao documento em relação a consulta.



Exercício 1

- Dado um sistema de RI de busca de livros, como o sistema de bibliotecas do IFMG.
  - Quem são os possíveis D,Q,F,R(qi,dj)?

Taxonomia de modelos de RI

Caracterização de um modelo de RI

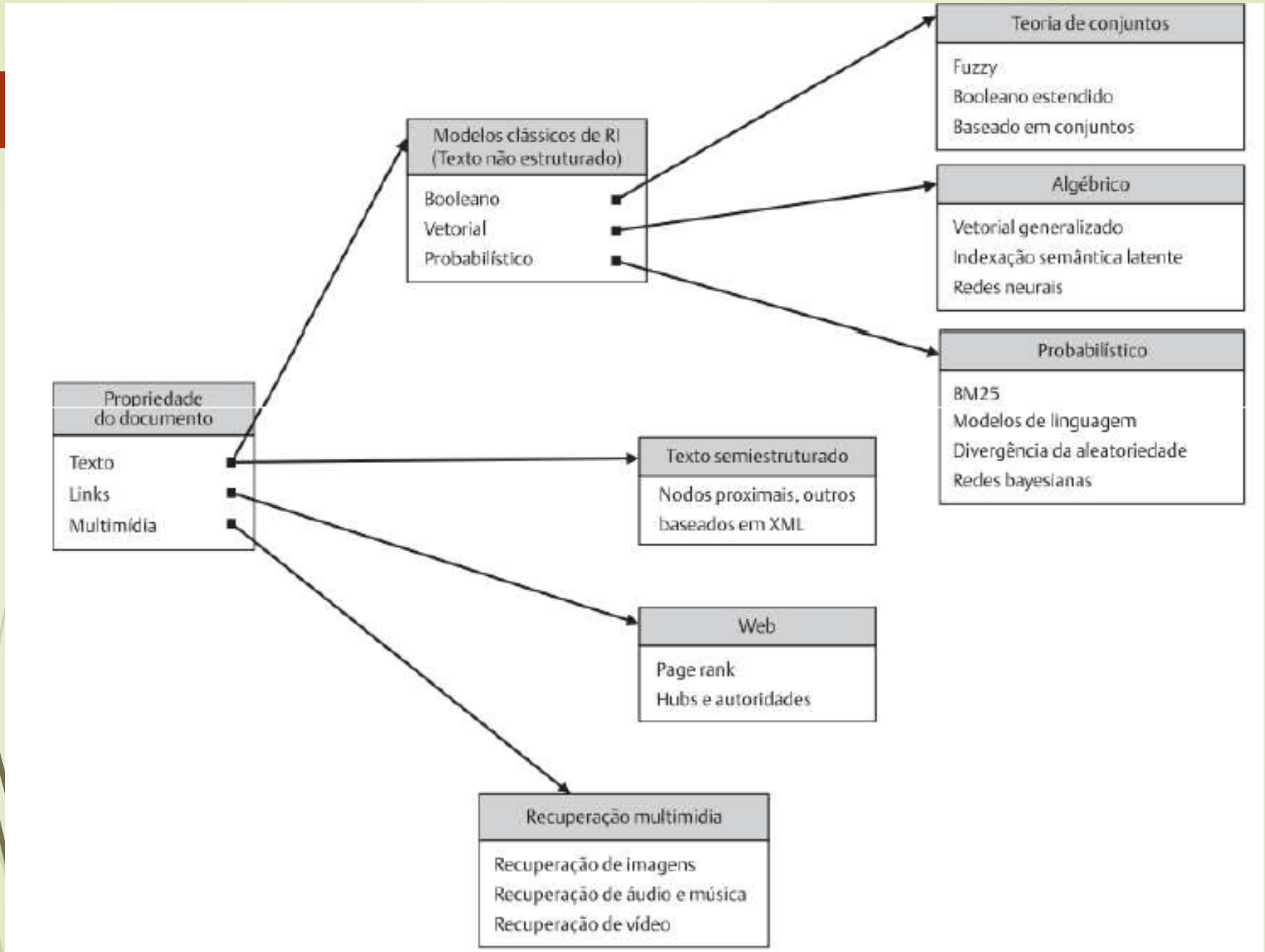
- **Modelos de RI são fundamentalmente baseados em texto**, isto é, eles usam o texto dos documentos para ranqueá-los em relação à consulta;

- **Na Web**, contudo, também é necessário utilizar a informação sobre a **estrutura de links** para alcançar um bom ranqueamento.

Caracterização de um modelo de RI

- **Objetos multimídia** (imagens e áudios) **não são codificados da mesma forma** que o texto e devem ser ranqueados de maneira diferente.

- Dadas essas características, distinguimos três categorias de modelos de RI:
  - Baseadas em texto;
  - Baseadas em links;
  - Baseadas em objetos multimídia.


Propriedade do documento
Texto
Links
Multimídia
Modelos clássicos de RI (Texto não estruturado)
Booleano
Vetorial
Probabilístico
Teoria de conjuntos
Fuzzy
Booleano estendido
Baseado em conjuntos
Algébrico
Vetorial generalizado
Indexação semântica latente
Redes neurais
Probabilístico
BM25
Modelos de linguagem
Divergência da aleatoriedade
Redes bayesianas
Texto semiestruturado
Nodos proximais, outros
baseados em XML
Web
Page rank
Hubs e autoridades
Recuperação multimídia
Recuperação de imagens
Recuperação de áudio e música
Recuperação de vídeo

Recuperação de informação clássica

Conceitos básicos

- Os modelos clássicos de RI consideram que cada documento é descrito por um conjunto de palavras-chave representativas, chamadas de **termos de indexação**:
  - uma palavra ou um grupo de palavras pré-selecionadas que representam conceitos-chave em um documento.

Conceitos básicos

- Um conjunto pré-selecionado de termos de indexação pode ser utilizado, por exemplo, para sumarizar o conteúdo dos documentos.

- Nesse caso, os termos são principalmente **substantivos** ou **grupos de substantivos**, uma vez que substantivos possuem significado próprio;

  - Adjetivos, advérbios e conectores são menos úteis como termos de indexação, pois funcionam principalmente como complementos.

Conceitos básicos - Vocabulário

- **Definição:**

- *Considere t como o número de termos de indexação na coleção de documentos e k_i como um termo de indexação genérico.*
- *$V = \{k_1, \ldots, k_t\}$ é o conjunto de todos os termos de indexação distintos na coleção e é comumente chamado de **vocabulário** V da coleção.*
  - *O tamanho do vocabulário é t.*

Conceitos básicos – Representação de documento e consulta

- **Definição:**
- *Considere V = {k_1, k_2, . . . , k_t} como o vocabulário da coleção. Se três termos de indexação k_l, k_m e k_n ocorrem em um mesmo documento d_j, dizemos que o padrão [k_l, k_m, k_n] de* ***coocorrência*** *de termos foi observado.*
  - *Cada um desses padrões de coocorrências de termos é chamado de componente* ***conjuntivo*** *de termo.*
- Exemplo: o padrão (1,0,...,0) indica a presença do termo k_1. O padrão (1,1,...,1) indica a presença de todos os termos no referido documento.

Conceitos básicos – Representação de documento e consulta

- Nesse caso, consultas e documentos são representados simplesmente pelos componentes conjuntivos de termo;

- Essa é a representação mais simples possível e é frequentemente conhecida como *bag of words* (saco de palavras).

Conceitos básicos – Representação de documento e consulta

- **A matriz de termos e documentos:**

- Um **outra forma de representar documento** é pela matriz de termos e documentos.
  - A ocorrência entre um termo em um documento estabelece uma relação entre eles.

- Essas **relações podem ser quantificadas**, por exemplo, pela frequência do termo no documento de forma matricial.

Conceitos básicos – Representação de documento e consulta

- **A matriz de**

$$\begin{array}{c} \\ k_1 \\ k_2 \\ k_3 \end{array} \begin{array}{c} \begin{array}{cc} d_1 & d_2 \end{array} \\ \begin{bmatrix} f_{1,1} & f_{1,2} \\ f_{2,1} & f_{2,2} \\ f_{3,1} & f_{3,2} \end{bmatrix} \end{array}$$

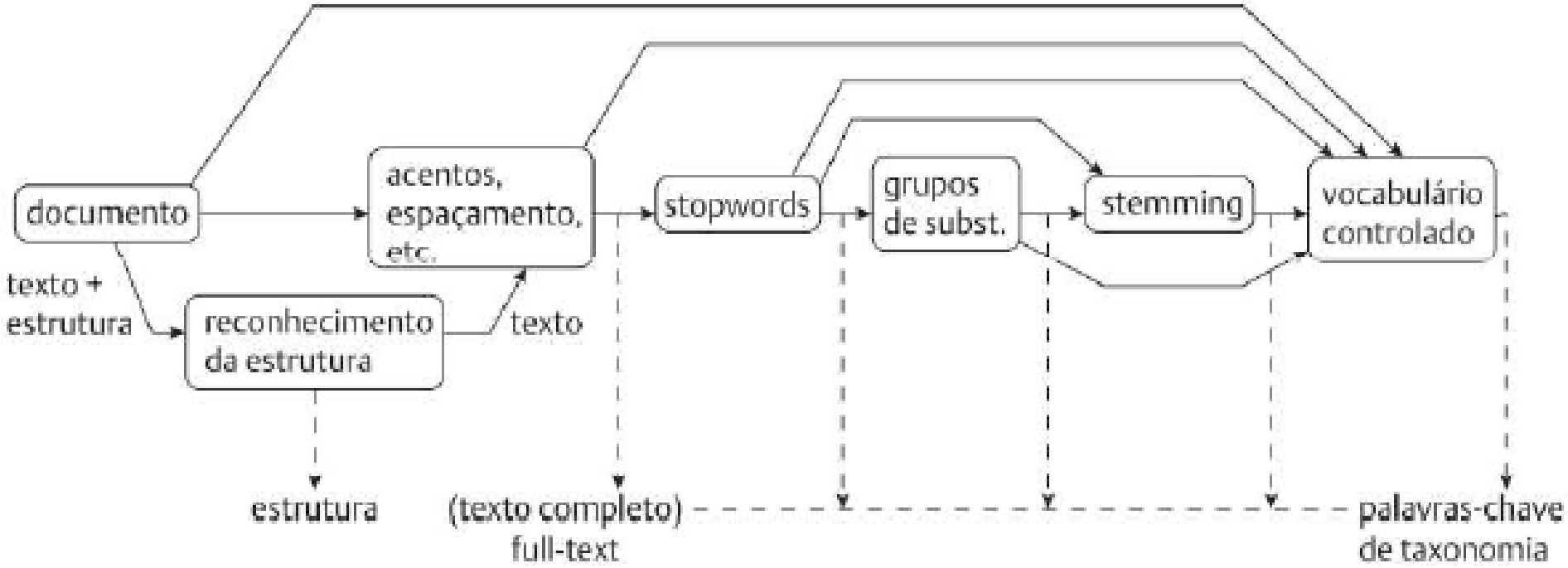
- Onde cada elemento $f_{i,j}$ representa a frequência do termo k_i no documento d_j.

- **Fornece mais informações** do que simples registrar se o termo aparece ou não no documento.
  - Mas ainda é uma abordagem simplista.

Conceitos básicos – Representação de documento

- Cada documento do conjunto pode ser representado por:
  - 1) Um conjunto de termos indexados que melhor representem seus tópicos:
    - Eliminar stopwords (artigos e preposições)
    - Aplicar stemming (reduz palavras distintas a sua raiz gramatical comum)
    - Eliminar adjetivos, advérbios e verbos
    - Veremos mais adiante;
  - 2) Texto completo;
  - 3) Texto completo + estrutura interna (capítulos, seções e etc).

Conceitos básicos – Representação de documento



Várias formas de representação de um documento por um sistema de RI.

Exercício 2

- Considere k_1 = {as} , k_2 = {um} e k_3 = {o}. Seja d_1 = {primeira estrofe do hino nacional} e d_2 = {segunda estrofe do hino nacional}.

- Monte a matriz de termos e documentos para esse sistema.

Exercício 2

- D1 = {Ouviram do Ipiranga as margens plácidas De um povo heroico o brado retumbante, E o sol da Liberdade, em raios fúlgidos, Brilhou no céu da Pátria nesse instante.}

- D2 = {Se o penhor dessa igualdade Conseguimos conquistar com braço forte, Em teu seio, ó Liberdade, Desafia o nosso peito a própria morte!}


Ótima segunda pra você!
Que seu começo de semana seja o início de muitas coisas boas também!
mensagensdebomdia.com.br