



INSTITUTO FEDERAL
MINAS GERAIS

RECUPERAÇÃO DE INFORMAÇÃO

Profa. Patrícia Proença
patricia.proenca@ifmg.edu.br

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

Introdução à Recuperação de Informação

Profa. Patrícia Aparecida Proença Avila

Roteiro

- Recuperação de informação (RI);
- Breve histórico;
- O problema de RI;
- O sistema de RI;

Introdução



- **Para pensarmos:**
 - **Dado x Informação**
- **Dado:** correspondem a um atributo, uma característica, uma propriedade de um objeto que sozinho, sem um contexto, não tem significado.
- **Informação:** é um conjunto de fatos organizados a terem valores significativos e úteis.

Introdução

- **Para pensarmos:**

- **Recuperação**

- **Pesquisa:**

- **Busca:**

Informação




Introdução



- A Recuperação de Informação (RI) é uma área abrangente da Ciência da Computação que estuda:
 - **representação, o armazenamento e acesso a itens de dados** (textos, imagens, vídeos, etc);
 - com o **objetivo de facilitar a tarefa do usuário de encontrar informação** de seu interesse localizadas em grandes coleções.

Introdução



- Outra Definição:
- 
- É o nome do **processo** onde um possível usuário de informação pode **converter** a sua **necessidade de informação** em uma lista real de **citações de documentos** armazenados que contenham informações **úteis** a ele...

Objetivos Iniciais (Históricos)

- Indexação de textos e busca por documentos úteis em uma coleção;
- Gerenciamento de acervos e bibliotecas;

Objetivos Atuais

- Classificação de textos;
 - Arquitetura de sistemas;
 - Interfaces de usuário;
 - Visualização de dados;
 - Filtros e linguagens.
-
- Exemplo: buscadores de RI modernos como o Google, Yahoo e outros.

Objetivos Atuais

- A área pode ser estudada sob dois pontos de vista distintos e complementares:
 - Centrado no computador;
 - Centrado no usuário

Objetivos Atuais

- **RI – Centrada no computador:**
- Consiste, principalmente, na construção de:
 - Índices eficientes;
 - Processamento de consultas com alto desempenho;
 - Desenvolvimento de novos algoritmos de ranqueamento, a fim de melhorar os resultados.

Objetivos Atuais

- **RI – Centrada no usuário:**
- Consiste, principalmente, em estudar:
 - O comportamento do usuário;
 - Entender suas principais necessidades;
 - Determinar como esse entendimento afeta a organização e a operação do sistema de recuperação.

Breve Histórico

Breve histórico da área de RI

- Por mais de 5000 anos, a humanidade vem organizando a informação para posterior busca e recuperação.
 - Tabuletas de argila, hieróglifos, rolos de papiros e livros
 - Para armazenar são usadas as bibliotecas.
- Considerando que o volume de informações nas bibliotecas está sempre crescendo é necessário construir estruturas de dados especializadas que possibilitem uma busca rápida – os **índices**.
 - Fornecem acesso rápido aos dados e aceleram o processamento das consultas.

Breve histórico da área de RI

- Durante séculos os índices foram criados manualmente como conjuntos de categorias.
- Compostas por rótulos, que identificam seus tópicos associados, e por ponteiros, para os documentos que discutem tais tópicos.
- Inicialmente criados por bibliotecários e hoje pelo computador;

Breve histórico da área de RI

- As bibliotecas estão entre as primeiras instituições a adotarem sistemas de RI para recuperar informações.
 - Primeira geração: automação de processos existentes, como a busca em catálogos de fichas, restritas ao nome do autor e ao título da obra.
 - Segunda geração: novas funcionalidades de busca foram adicionadas para incluir assuntos, palavras-chave e operadores de consulta.
 - Terceira Geração: foco em interfaces gráficas melhoradas, formulários eletrônicos, entre outros.

Breve histórico da área de RI

- Apesar de sua maturidade, até recentemente a RI era vista como uma área de interesse limitada apenas a bibliotecários e a especialistas em informação;
- O que alterou esse interesse?

Breve histórico da área de RI

- O surgimento da Web!
- A “grande rede” tornou-se um repositório universal da cultura e do conhecimento humano;
- Devido a facilidade de acesso e uso, milhões de usuários criaram bilhões de documentos que compõem o maior repositório humano de conhecimento na história.
- Como encontrar informações úteis na Web??

O problema de RI

Diferentes necessidades de informação

- **Usuários de sistemas modernos de RI**, como usuários de máquinas de busca, **têm necessidades de informação de diferentes níveis de complexidade:**
 - No caso mais simples, eles procuram pelo **link** para a página de uma empresa, governo ou instituição;
 - Nos casos mais sofisticados, procuram por **informações necessárias** à execução de uma tarefa associada a seus trabalhos ou a necessidades imediatas.

Necessidade de informação complexa

- Um exemplo de uma necessidade de informação mais complexa é:
 - Encontre todos os documentos que tratam do papel do governo federal no financiamento das operações da Petrobrás.
- Essa **descrição completa da necessidade** do usuário **não necessariamente** fornece a **melhor formulação de consulta** para o sistema de RI.

O problema de RI

- Dada a consulta do usuário, o objetivo maior do sistema de RI é recuperar informações que sejam úteis ou relevantes para o usuário;
- A ênfase está na **recuperação de informação**, não na **recuperação de dados**.

O problema de RI

- O sistema de RI deve de **alguma forma “interpretar” o conteúdo dos itens** de informação de uma coleção:
- **classificá-los de acordo com o grau de relevância à consulta do usuário;**

O problema de RI



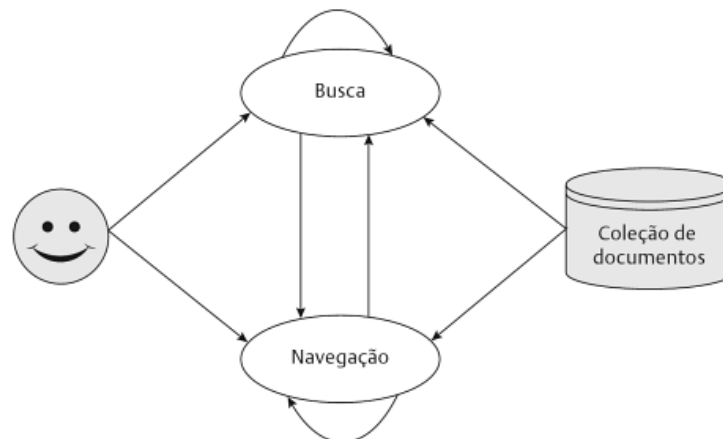
- “O objetivo principal de um sistema de RI é **recuperar todos os documentos que são relevantes à necessidade de informação do usuário e, ao mesmo tempo, recuperar o menor número possível de documentos irrelevantes.**”

O problema de RI

- Dificuldades:
 - 1) Como extrair informações dos documentos?
 - 2) Como utilizar tais informações para decidir sobre a sua relevância?
 - Relevância é um julgamento pessoal que depende da tarefa a ser resolvida e de seu contexto.
 - Relevância é algo subjetivo e pode mudar de acordo com o tempo e local.
- Nesse sentido, nenhum sistema de RI pode fornecer respostas perfeitas a todos os usuários o tempo todo.

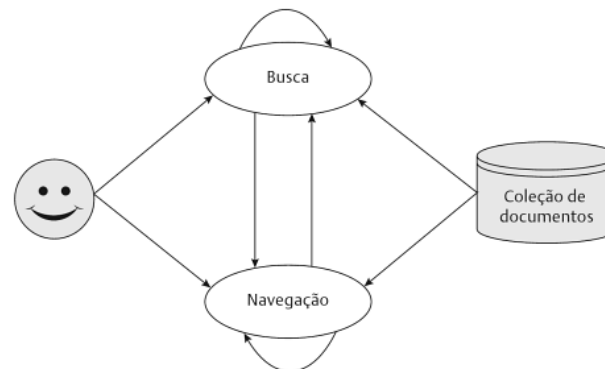
Tarefa do usuário

- O usuário de um sistema de RI precisa traduzir sua necessidade de informação em uma consulta na linguagem fornecida pelo sistema.
- Geralmente um conjunto de palavras ou imagens;
- Dizemos que o usuário está buscando ou consultando informações de seu interesse.



Tarefa do usuário

- Quando o usuário possui um interesse que não está bem definido ou que é muito amplo dizemos que ele está navegando pelos documentos da coleção e não buscando.
- Exemplo documentos sobre corridas de automóveis em geral e decide olhar documentos sobre Fórmula 1 e Fórmula Indy.



RI x Recuperação de dados

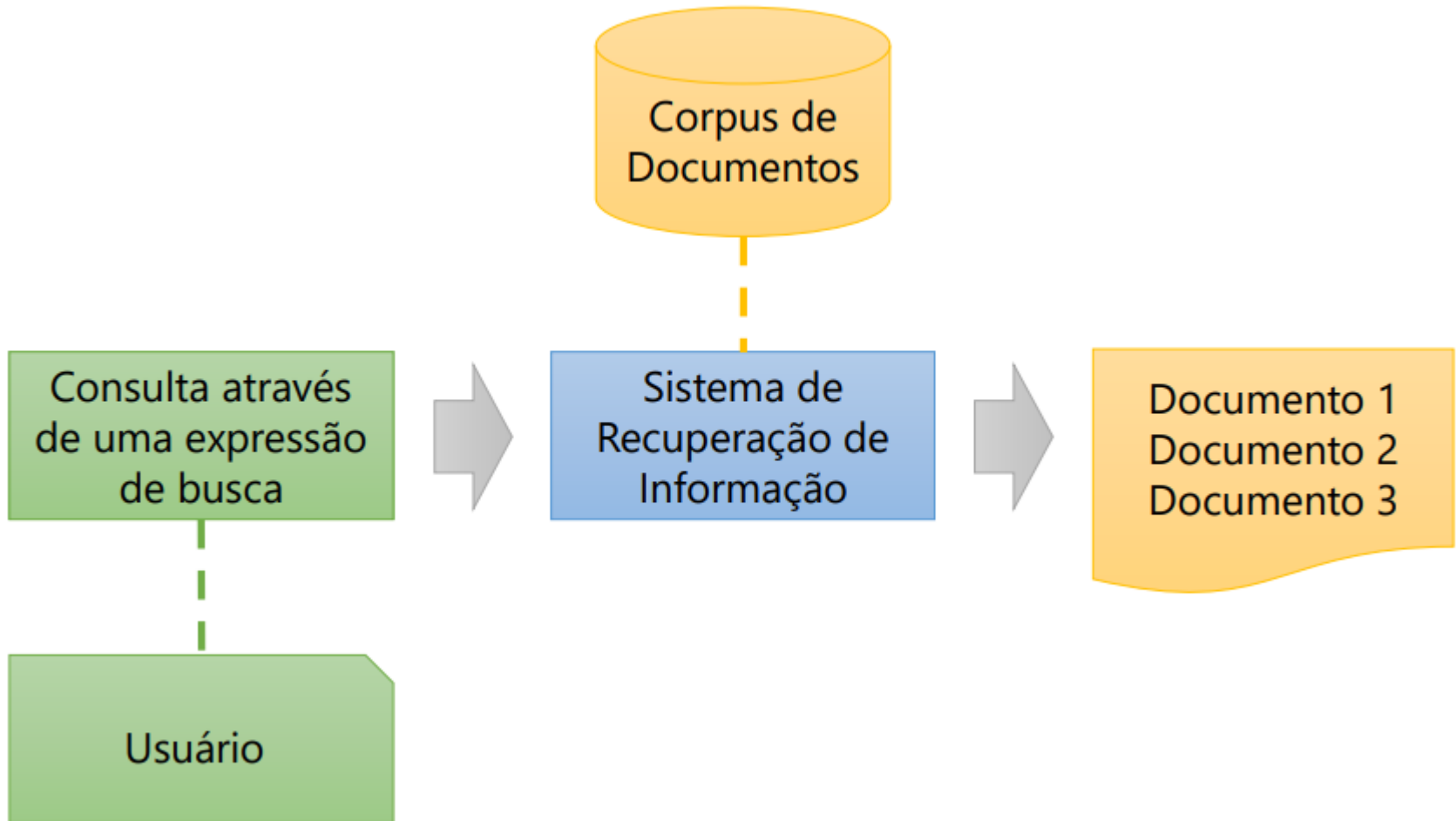
- A recuperação de dados, no contexto de um sistema de RI, consiste na identificação de **quais documentos da coleção contêm as palavras-chave da consulta do usuário**;
- Com frequência, isso não é suficiente para satisfazer as necessidades de informação do usuário;

RI x Recuperação de dados

- O usuário de um sistema de RI está mais interessado em **recuperar informações** sobre um assunto do que em **recuperar dados** que satisfaçam uma dada consulta.
 - Por exemplo, um usuário de um sistema RI está disposto a aceitar, no conjunto dos resultados, documentos que possuam sinônimos dos termos usados na consulta.

O Sistema de RI

Elementos Básicos



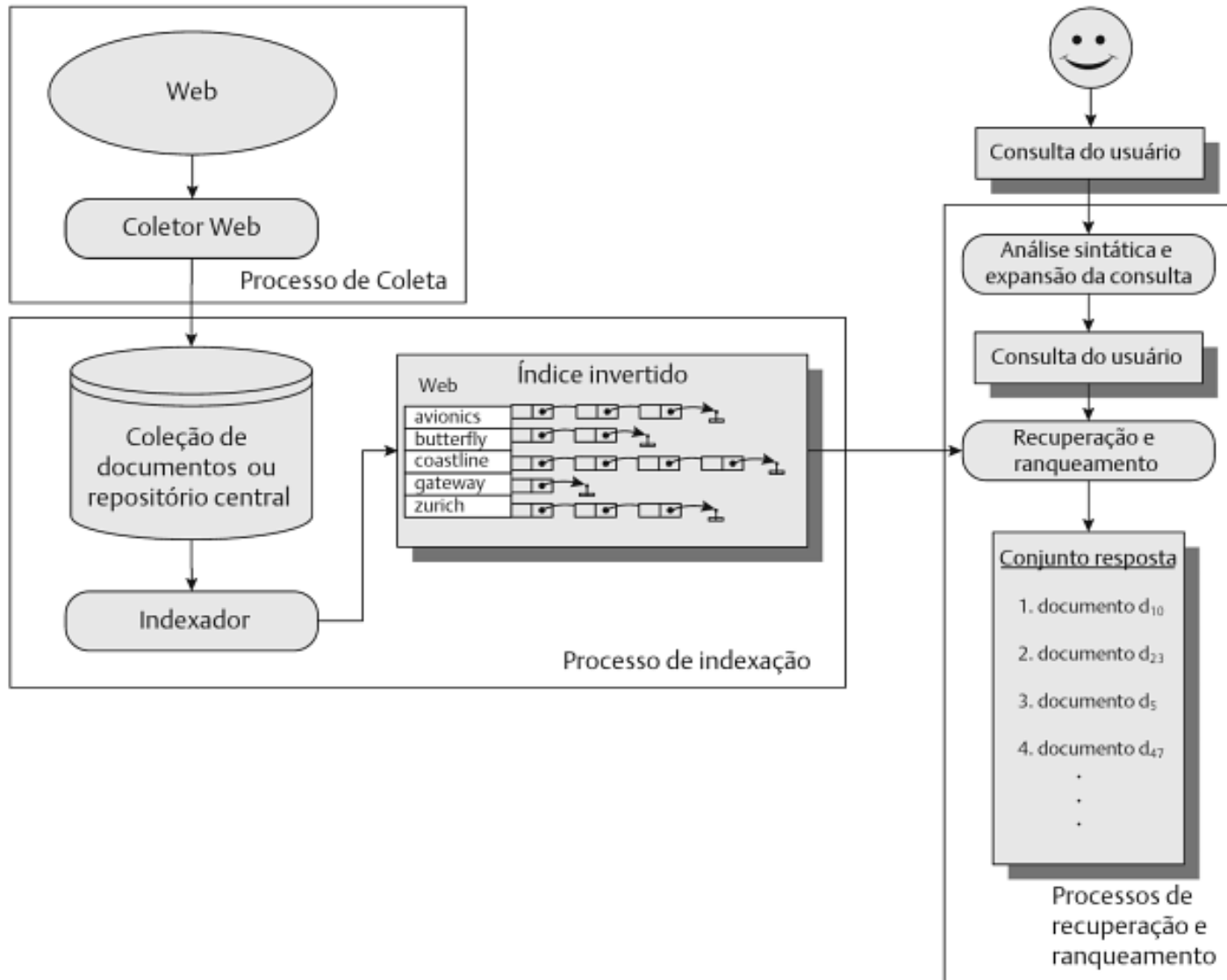
O sistema de RI

- A fim de descrever o sistema de RI, utilizaremos uma arquitetura de software simples e genérica composta por seis módulos:
 - 1) Obtenção da coleção de documentos;
 - Que pode ser particular ou coletada na web;
 - 2) Indexação dos documentos;
 - Os documentos armazenados são indexados para facilitar a recuperação e o ranqueamento;
 - A estrutura de índice mais utilizada é o índice invertido;
 - 3) Consulta do usuário;
 - A consulta é analisada sintaticamente e expandida, por exemplo, formas variantes das palavras;

O sistema de RI

- 4) Recuperação de documentos;
 - A consulta do sistema é processada, utilizando-se o índice para recuperar um subconjunto dos documentos.
- 5) Ranqueamento dos documentos;
 - Identificar os documentos que tem maior probabilidade de serem considerados relevantes – parte mais critica de um sistema de RI.
- 6) Apresentação para o usuário.
 - Documentos que estão no topo do ranking são retornados ao usuário.

O sistema de RI



Realização de uma busca em um sistema de RI

- Passos:
- 1) Usuário especifica uma consulta que reflete sua necessidade de informação;
- 2) A consulta é analisada sintaticamente e expandida;
- 3) A consulta expandida (ou do sistema) é então processada utilizando-se o índice para recuperar um subconjunto dos documentos;
- 4) Os documentos recuperados são ranqueados e aqueles que estão no topo do ranking são apresentados ao usuário (parte mais crítica de um sistema de RI).

Recapturando

No decorrer da aula vimos...

- O objetivo da área de estudo conhecida como Recuperação de Informação;
 - 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;

No decorrer da aula vimos...

- 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..

Próximas aulas

- Estudo dos modelos clássicos de recuperação e ranqueamento de documentos:
 - Modelo booleano;
 - Modelo vetorial;
 - Modelo probabilístico.

Atividade para entregar – 18/05

- Objetivo:
 - Fazer um programa simples que busca em um arquivo longo por uma string que o usuário deseja procurar.
- Entradas esperadas
 - O programa deverá receber como entrada um nome do arquivo a ser processado, uma palavra procurada e um nome de arquivo de saída a ser escrito.

Atividade para entregar – 18/05

- Saídas produzidas:
 - Programa deve mostrar em um arquivo de saída o número de vezes que a string procurada aparece e escrever as linhas que contêm a string alvo no arquivo de saída.
- Exemplo:
 - 120 /*número de vezes que a palavra foi encontrada.
 - 1
 - 2
 - 50
 -

Atividade para entregar – 18/05

- Testar com os arquivos disponíveis na descrição da atividade:
 - Teste.zip
 - A palavra “informação”
 - O número de ocorrências deve ser 11 e o número de linhas no arquivo de saída deve ser 11.
 - Corpus.zip
 - A palavra “Brazil”.
 - O número de ocorrências deve ser 1281;



Mantenha o foco no
objetivo, centralize a
força para lutar e utilize
a fé para vencer.