Organização e Recuperação de Informação – GSI024

Prof. Dr. Rodrigo Sanches Miani – FACOM/UFU

Introdução

Organização e Recuperação de Informação (GSI024)

Tópicos

- Recuperação de informação (RI);
- Breve histórico;
- O problema de RI;
- O sistema de RI;
- Impacto da Web nos sistemas de RI.

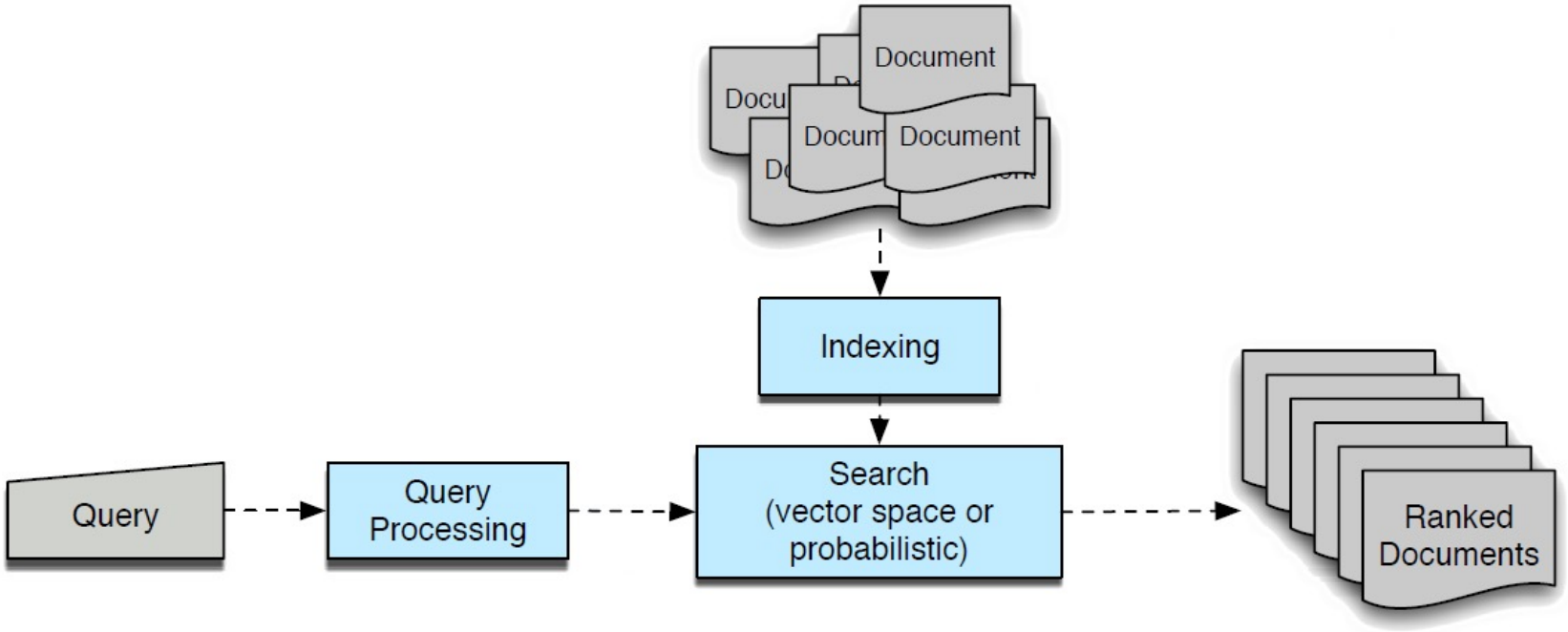
Recuperação de informação

Organização e Recuperação de Informação (GSI024)

Recuperação da Informação (RI)

- A Recuperação de Informação (RI) é uma área abrangente da Ciência da Computação que se concentra principalmente em prover aos usuários o acesso fácil às informações de seu interesse.

Recuperação da Informação (RI)



Extraído de https://devopedia.org/information-retrieval

RI – Uma possível definição

- *A Recuperação de Informação trata da **representação**, **armazenamento**, **organização** e **acesso** a itens de informação, como documentos, páginas Web, catálogos online, registros estruturados e semiestruturados, objetos multimídia, etc. A representação e a organização dos itens de informação devem fornecer aos usuários facilidade de acesso às informações de seu interesse.* (Baeza-Yates, Ribeiro-Neto, 2013)

RI – Objetivos iniciais

- Indexação de textos e busca por documentos úteis em uma coleção;
- Gerenciamento de acervos e bibliotecas;
- Exemplo: construção de índices para a busca eficiente de informações em bibliotecas.

RI – Objetivos atuais

- Classificação de textos;
- Arquitetura de sistemas;
- Interfaces de usuário;
- Visualização de dados;
- Filtros e linguagens.

RI – Objetivos atuais

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

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.

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

Organização e Recuperação de Informação (GSI024)

Breve histórico da área de RI

- Desenvolvimentos iniciais na área de RI foram realizados nos anos 50 por meio dos esforços de pesquisa dos seguintes pesquisadores: Hans Peter Luhn, Eugene Garfield, Philip Bagley e Calvin Moores;

- De acordo com diversas referências, Calvin Moores foi o responsável por cunhar o termo "*Information Retrieval*".

Breve histórico da área de RI

- **1952:** H.P. Luhn propõe, usando cartões perfurados, uma das primeiras implementações do modelo booleano;
- **1955**: Allen Kent e seus colegas publicaram um artigo descrevendo as métricas de precisão e revocação;
- 1957: H.P. Luhn propõe uma abordagem estatística para o problema de RI;
- **1962**: Paradigma de Cranfield por Cleverdon;
- **1963**: Joseph Becker e Robert Hayes publicaram o primeiro livro sobre RI;
- **1968**: Publicação do primeiro livro de RI por Gerard Salton;
- **1978**: primeiro congresso da ACM sobre RI (ACM SIGIR) acontece em Rochester, Nova York;
- **1983**: Salton e McGill publicaram Introduction to Modern Information Retrieval, um livro clássico em RI focado no modelo vetorial;
- **1998**: Google;
- **Meados de 2010**: Aplicação de redes neurais em Recuperação da Informação.

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!
  - Web e Internet?
- A "grande rede" é um repositório universal da cultura e do conhecimento humano;
- Como encontrar informações úteis na Web??

O problema de RI

Organização e Recuperação de Informação (GSI024)

Diferentes necessidades de informação

Usuários de sistemas modernos de RI (usuários de máquinas de busca na Web, por exemplo), têm necessidades de informação de diferentes níveis de complexidade:

1. No caso mais simples, eles procuram pelo **link** para a página de uma empresa, governo ou instituição;
2. 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

- O usuário pode querer primeiro traduzir essa necessidade de informação em uma consulta ou em uma sequência de consultas a serem submetidas ao sistema;

- Essa tradução gera uma série de palavras-chave, ou **termos de indexação**, que sumarizam a necessidade de informação do usuário.

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**.

- Em breve veremos a diferença entre esses dois conceitos...

O problema de RI

- O sistema de RI deve, de alguma forma, "interpretar" o conteúdo dos itens de informação:
  - Dentre os documentos de uma coleção, classifica-los de acordo com o grau de relevância à consulta do usuário;

Como?

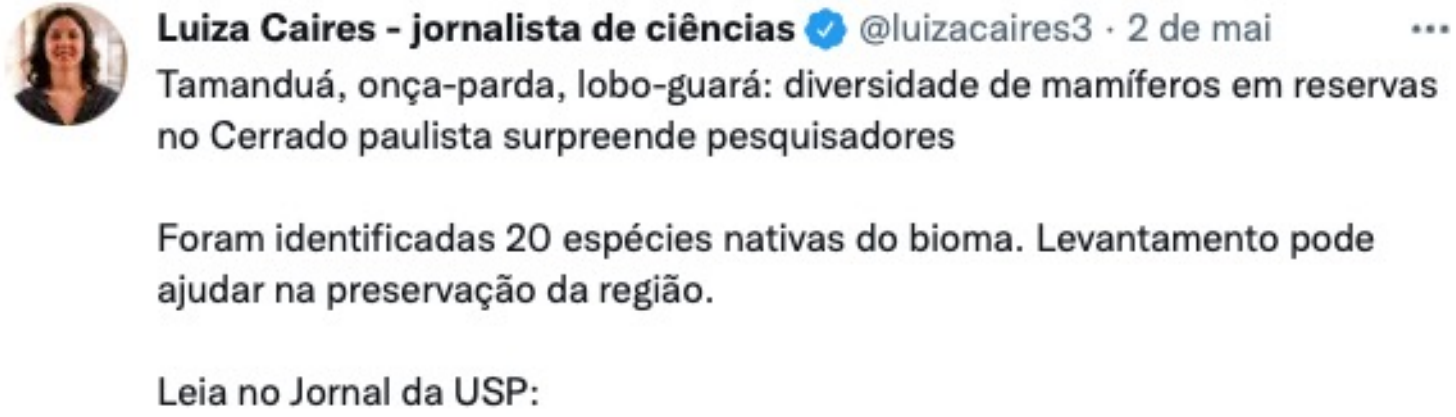
- Essa "interpretação" do conteúdo de um documento envolve a extração de informações sintáticas e semânticas do texto do documento e sua utilização para satisfazer a necessidade de informação do usuário.

O problema de RI - Exemplo

Atila Iamarino *ainda de licença paternidade @oatila · 23 h

O H1N1 já tinha circulado entre humanos antes. Faz sentido o SARS-CoV-2, que mal entrou em humanos, ter ainda mais espaço para melhorar sua transmissão. O grande problema é que isso facilita ondas ainda mais bruscas do que a Omicron em Dez/Jan e dificulta o trabalho das vacinas.

9 13 250



Luiza Caires - jornalista de ciências @luizacaires3 · 2 de mai

Tamanduá, onça-parda, lobo-guará: diversidade de mamíferos em reservas no Cerrado paulista surpreende pesquisadores

Foram identificadas 20 espécies nativas do bioma. Levantamento pode ajudar na preservação da região.

Leia no Jornal da USP:

O problema de RI - Exemplo

- O ser humano, ao bater o olho, nos tweets sabe diferenciar o conteúdo/assunto de ambos os documentos.
- O que poderíamos fazer para que o computador também possa diferenciar tais tweets?
- Ou seja, se a minha consulta estiver relacionada sobre "vacinas covid", o tweet do Átila deveria ser melhor ranqueado do que o tweet da Luiza.

O problema de RI

O problema da 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 algo subjetivo e pode mudar de acordo com o tempo, local ou até mesmo de acordo com o dispositivo.

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.

Pergunta: Diferença entre dado e informação?

RI x Recuperação de dados

Uma distinção fundamental entre dado e informação é que o primeiro é puramente *sintático* e a segunda contém necessariamente *semântica*.

Dados podem ser totalmente descritos através de representações formais, estruturais enquanto que a informação é uma abstração informal.

RI x Recuperação de dados - Exemplo

Dado

- Um Macbook Pro de 13 polegadas custa R$ 17.299,00.
- A temperatura atual de Bauru-SP é de 16 graus.

Informação

- Um Macbook Pro de 13 polegadas custa, em média, 45% mais caro do que os notebooks vendidos pela Dell.
- A média da temperatura da cidade de Bauru-SP no período da noite e no mês de dezembro é 23 graus, portanto, é possível dizer que hoje é um dia mais ameno do que o normal.

RI x Recuperação de dados

O processo de recuperação de informação consiste em identificar, no conjunto de documentos do sistema, quais atendem à necessidade de **informação** do usuário;

- Na maioria das vezes, o usuário está interessado em recuperar informação sobre um determinado assunto e não em recuperar registros de dados.

RI x Recuperação de dados

Um **sistema de recuperação de dados**, como um banco de dados relacional, trata de dados que possuem estrutura semântica bem definidas (linguagem SQL, por exemplo);

Um **sistema de RI** lida com texto em linguagem natural que não é bem estruturado (buscas realizadas no Google, por exemplo).

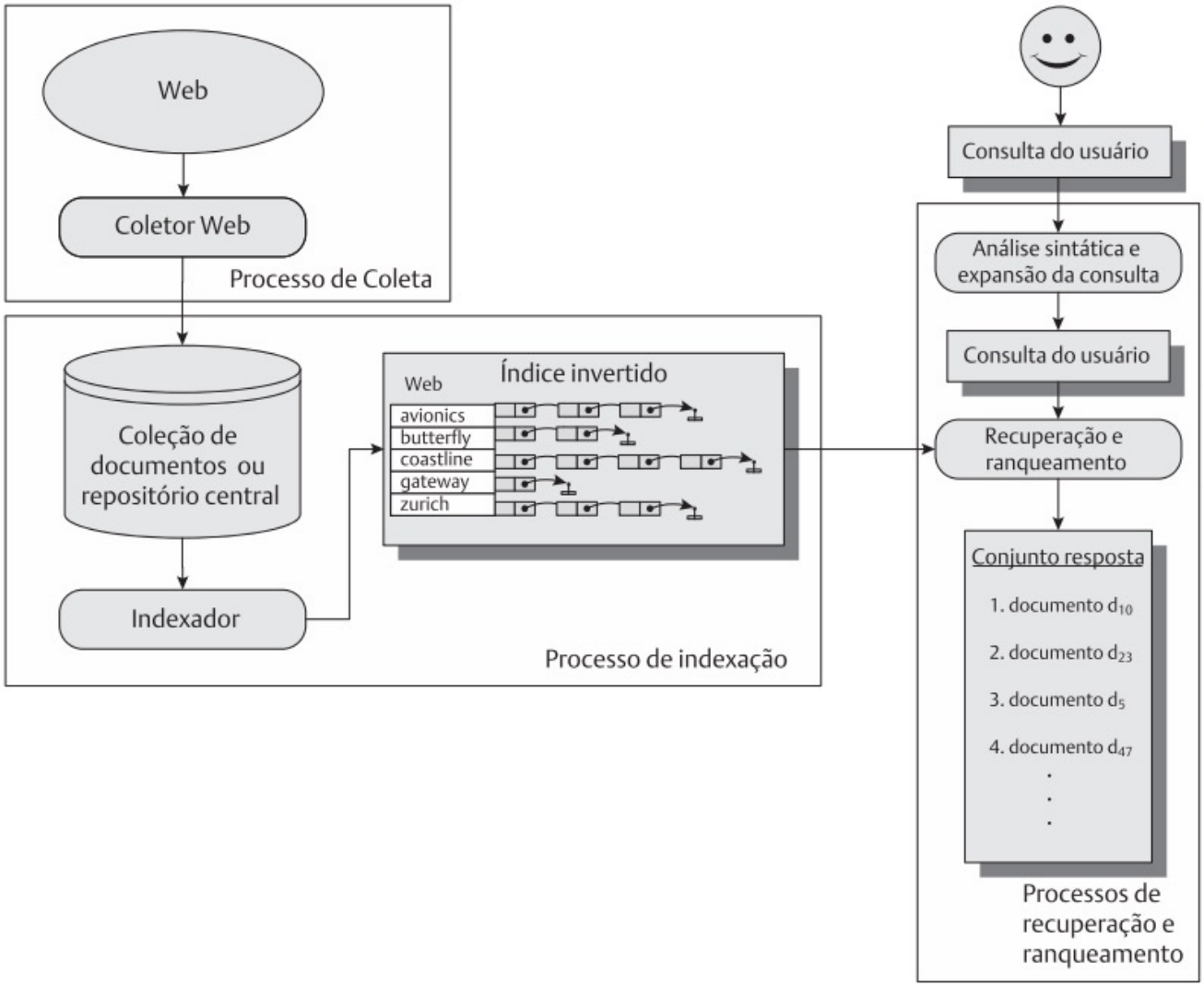
O Sistema de RI

Organização e Recuperação de Informação (GSI024)

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 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.

O sistema de RI



Realização de uma busca em um sistema de RI

- É bom acostumar com os diversos termos da área:
  1) Consulta;
  2) Recuperação;
  3) Ranqueamento;
  4) Conjunto resposta;
  5) Indexação;
  6) Coleção de documentos;

Realização de uma busca em um sistema de RI

- Para realizar uma busca, os seguintes passos são realizados:
  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).

Relevância de um documento

- Dado que a tarefa de decidir quanto à relevância de um documento é inerentemente subjetiva, avaliar a **qualidade do conjunto-resposta** é a chave para a melhoria do sistema de RI;

- Um processo de avaliação sistemático permite refinar o algoritmo de ranqueamento e **melhorar a qualidade dos resultados**;

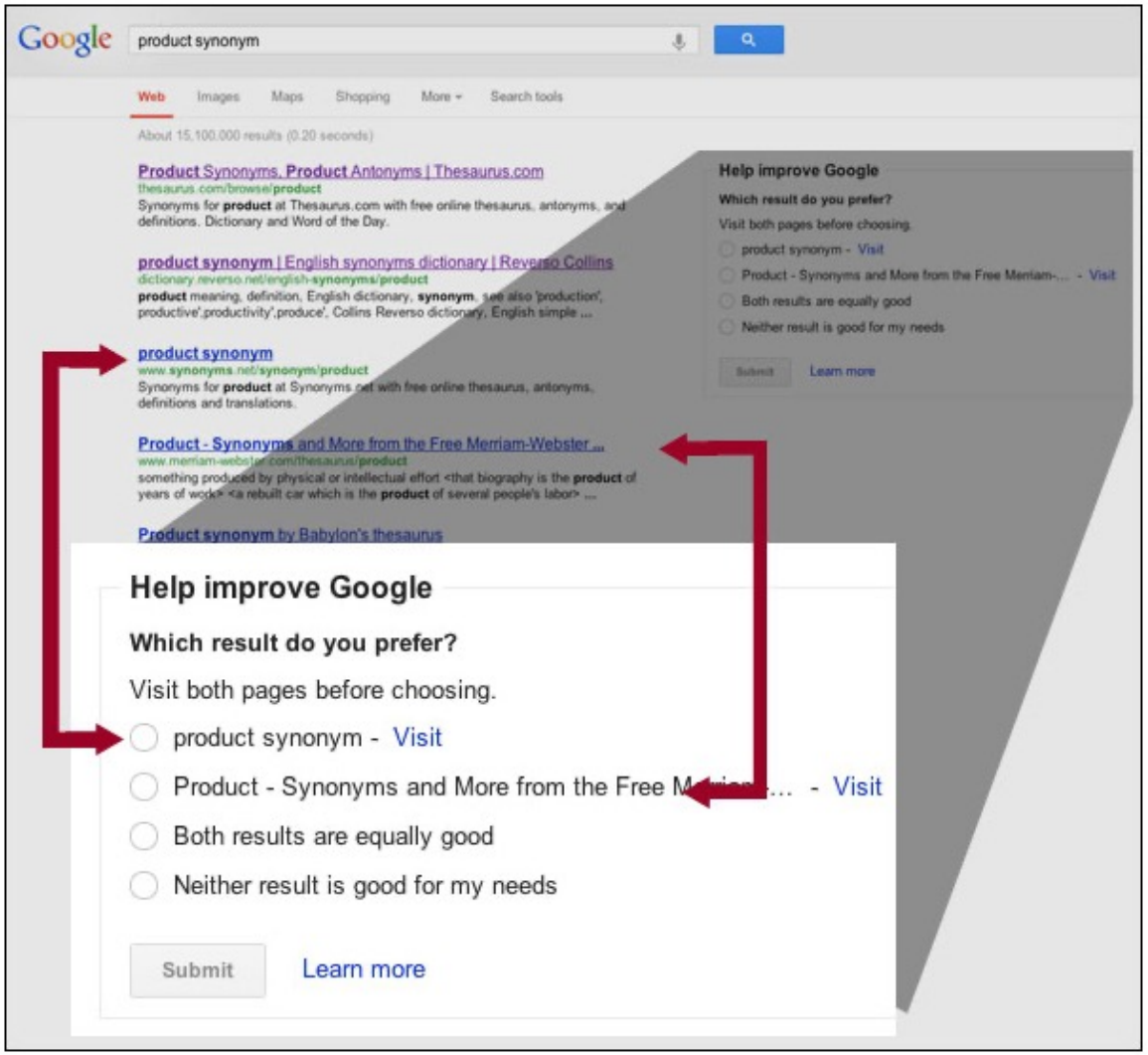
- O procedimento de avaliação mais comum consiste em comparar o conjunto de resultados produzidos pelo sistema de RI com os resultados sugeridos pelos especialistas humanos.

Relevância de um documento - Google

Até o Google usa o feedback dos próprios usuários para melhorar os resultados de suas consultas...

- http://searchengineland.com/google-feedback-experiment-which-result-142872
- https://www.mattcutts.com/blog/the-role-of-humans-in-google-search/

Relevância de um documento - Google



Sistemas de RI – Processos de recuperação e ranqueamento

- A fim de descrever o sistema de RI, utilizaremos uma arquitetura de software simples e genérica composta por 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.

Sistemas de RI – Processo de recuperação

- Fase 1 (Representação dos documentos)
  - Aplicação de operações textuais como a eliminação de stopwords, radicalização (stemming) sobre os documentos;
  - Seleção de um subconjunto de termos para serem utilizados como termos de indexação;
    - Os termos de indexação são utilizados para compor a representação do documento, que pode ser menor do que o documento original (dependendo do subconjunto de termos de indexação selecionado).
- Fase 2 (Indexação)
  - Relacionar os termos com os documentos.

Sistemas de RI – Processo de ranqueamento

- Os documentos recuperados são ranqueados de acordo com a **probabilidade de relevância** para o usuário;
- Fase mais crítica, pois a **qualidade** do resultado percebida pelo usuário depende diretamente do **ranqueamento**;
- Os documentos no topo do ranking são então formatados e apresentados para o usuário.

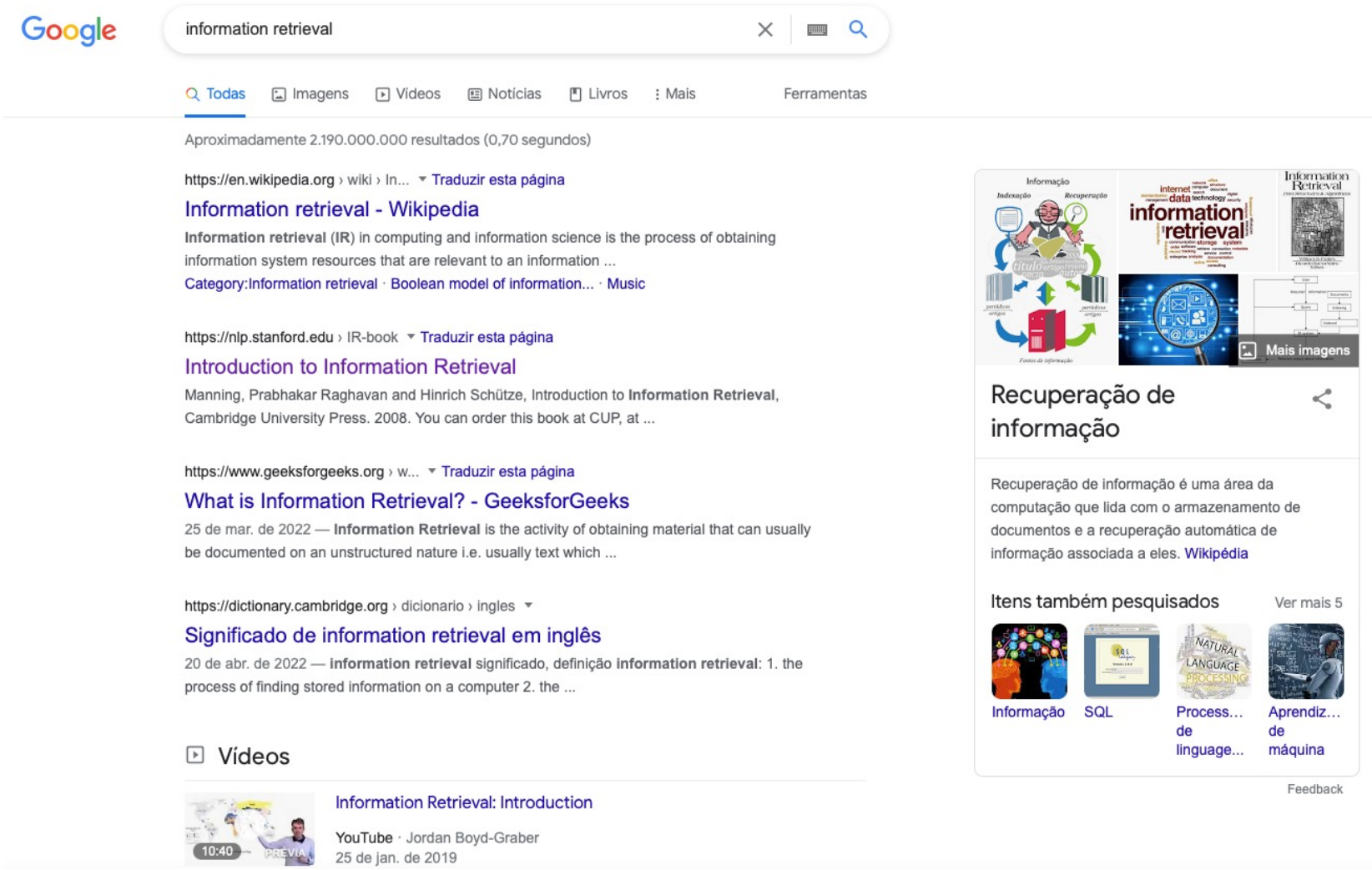
O impacto da Web em sistemas de RI

Organização e Recuperação de Informação (GSI521)

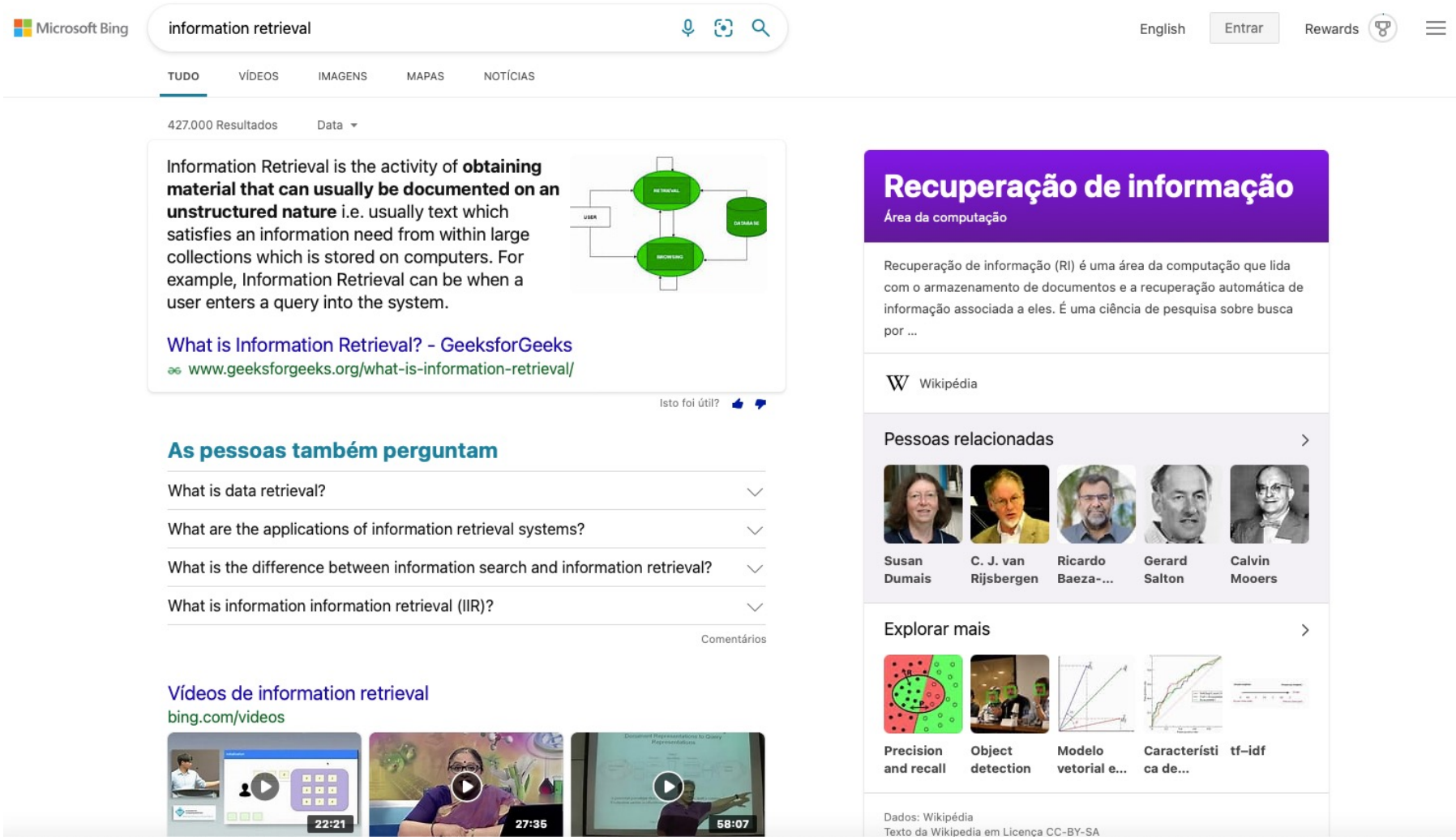
Busca na Web

- A busca na Web é a aplicação mais proeminente de RI e suas técnicas;
- Os componentes de recuperação e indexação de qualquer máquina de busca na Web são fundamentalmente tecnologias de RI;
- Uma consequência imediata é que a Web teve um grande impacto no desenvolvimento da RI.

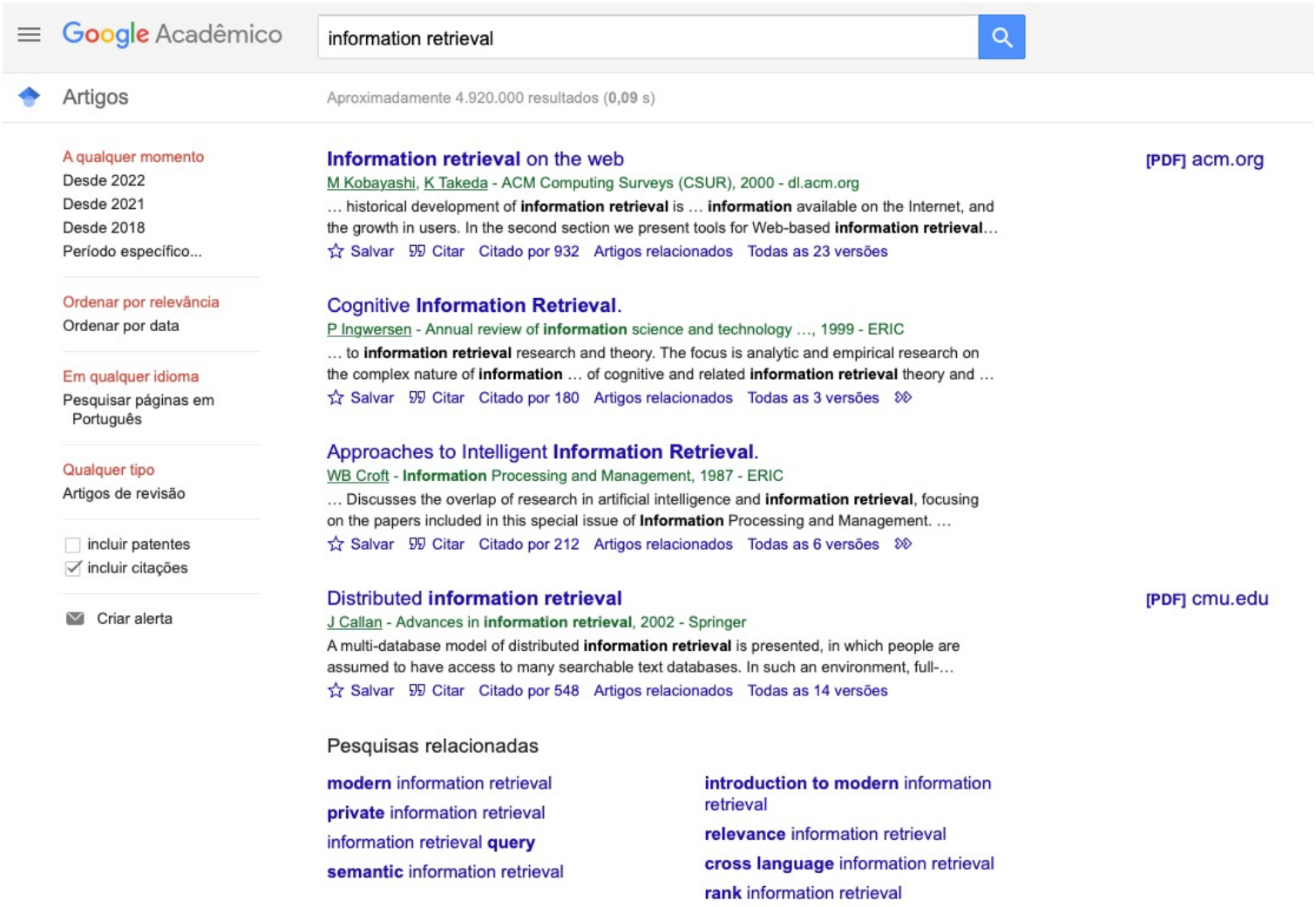
Busca na Web


Google
information retrieval
Todas
Imagens
Vídeos
Notícias
Livros
Mais
Ferramentas
Aproximadamente 2.190.000.000 resultados (0,70 segundos)
https://en.wikipedia.org › wiki › In... Traduzir esta página
Information retrieval - Wikipedia
Information retrieval (IR) in computing and information science is the process of obtaining information system resources that are relevant to an information ...
Category:Information retrieval · Boolean model of information... · Music
https://nlp.stanford.edu › IR-book Traduzir esta página
Introduction to Information Retrieval
Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008. You can order this book at CUP, at ...
https://www.geeksforgeeks.org › w... Traduzir esta página
What is Information Retrieval? - GeeksforGeeks
25 de mar. de 2022 — Information Retrieval is the activity of obtaining material that can usually be documented on an unstructured nature i.e. usually text which ...
https://dictionary.cambridge.org › dicionario › ingles
Significado de information retrieval em inglês
20 de abr. de 2022 — information retrieval significado, definição information retrieval: 1. the process of finding stored information on a computer 2. the ...
Vídeos
10:40
PRÉVIA
Information Retrieval: Introduction
YouTube · Jordan Boyd-Graber
25 de jan. de 2019
Mais imagens
Recuperação de informação
Recuperação de informação é uma área da computação que lida com o armazenamento de documentos e a recuperação automática de informação associada a eles. Wikipédia
Itens também pesquisados
Ver mais 5
Informação
SQL
Process… de linguage...
Aprendiz… de máquina
Feedback

Busca na Web


Microsoft Bing
information retrieval
English
Entrar
Rewards
TUDO
VÍDEOS
IMAGENS
MAPAS
NOTÍCIAS
427.000 Resultados
Data
Information Retrieval is the activity of **obtaining material that can usually be documented on an unstructured nature** i.e. usually text which satisfies an information need from within large collections which is stored on computers. For example, Information Retrieval can be when a user enters a query into the system.
What is Information Retrieval? - GeeksforGeeks
www.geeksforgeeks.org/what-is-information-retrieval/
Isto foi útil?
As pessoas também perguntam
What is data retrieval?
What are the applications of information retrieval systems?
What is the difference between information search and information retrieval?
What is information information retrieval (IIR)?
Comentários
Vídeos de information retrieval
bing.com/videos
22:21
27:35
58:07
Recuperação de informação
Área da computação
Recuperação de informação (RI) é uma área da computação que lida com o armazenamento de documentos e a recuperação automática de informação associada a eles. É uma ciência de pesquisa sobre busca por ...
Wikipédia
Pessoas relacionadas
Susan Dumais
C. J. van Rijsbergen
Ricardo Baeza-...
Gerard Salton
Calvin Mooers
Explorar mais
Precision and recall
Object detection
Modelo vetorial e...
Característi ca de...
tf–idf
Dados: Wikipédia
Texto da Wikipedia em Licença CC-BY-SA

Busca na Web


Google Acadêmico
information retrieval
Artigos
Aproximadamente 4.920.000 resultados (0,09 s)
A qualquer momento
Desde 2022
Desde 2021
Desde 2018
Período específico...
Ordenar por relevância
Ordenar por data
Em qualquer idioma
Pesquisar páginas em Português
Qualquer tipo
Artigos de revisão
incluir patentes
incluir citações
Criar alerta
Information retrieval on the web
M Kobayashi, K Takeda - ACM Computing Surveys (CSUR), 2000 - dl.acm.org
... historical development of information retrieval is ... information available on the Internet, and the growth in users. In the second section we present tools for Web-based information retrieval...
Salvar Citar Citado por 932 Artigos relacionados Todas as 23 versões
[PDF] acm.org
Cognitive Information Retrieval.
P Ingwersen - Annual review of information science and technology ..., 1999 - ERIC
... to information retrieval research and theory. The focus is analytic and empirical research on the complex nature of information ... of cognitive and related information retrieval theory and ...
Salvar Citar Citado por 180 Artigos relacionados Todas as 3 versões
Approaches to Intelligent Information Retrieval.
WB Croft - Information Processing and Management, 1987 - ERIC
... Discusses the overlap of research in artificial intelligence and information retrieval, focusing on the papers included in this special issue of Information Processing and Management. ...
Salvar Citar Citado por 212 Artigos relacionados Todas as 6 versões
Distributed information retrieval
J Callan - Advances in information retrieval, 2002 - Springer
A multi-database model of distributed information retrieval is presented, in which people are assumed to have access to many searchable text databases. In such an environment, full-...
Salvar Citar Citado por 548 Artigos relacionados Todas as 14 versões
[PDF] cmu.edu
Pesquisas relacionadas
modern information retrieval
private information retrieval
information retrieval query
semantic information retrieval
introduction to modern information retrieval
relevance information retrieval
cross language information retrieval
rank information retrieval

Impacto da Web no desenvolvimento de sistemas de RI

1) Características da coleção de documentos;
   - Documentos distribuídos conectados por hiperlinks.

2) Tamanho da coleção de documentos e o volume de consultas;
   - Desempenho e escalabilidade tornaram-se características importantes em sistemas de RI;
   - Em uma coleção muito grande, é difícil prever a relevância de documentos;

3) A Web não representa somente um repositório de documentos, mas também um meio de realização de negócios.
   - Pesquisa de preços, números de telefones, links para baixar softwares e etc.

Comentários

Organização e Recuperação de Informação (GSI521)

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;
- Breve histórico.

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

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

Estudos

- Recuperação de Informação: Conceitos e Tecnologia das Máquinas de Busca
  - Estudar o Capítulo 1

- Iniciar os estudos da linguagem Python. Sugestão:
  - https://www.coursera.org/learn/ciencia-computacao-python-conceitos

- Se cadastrar na plataforma Kaggle – dar uma olhada no seguinte conjunto de dados:
  - https://www.kaggle.com/datasets/augustop/portuguese-tweets-for-sentiment-analysis