



RECUPERAÇÃO DE INFORMAÇÃO (RI)

Emmanuel Silva Xavier



SOBRE MIM

Emmanuel Silva Xavier

Mestre em Engenharia da Computação
Especialista em Análise de Sistemas
Graduado em Sistemas de Informação
Professor de Ciências da Computação e
Engenharia Elétrica no IFMA



SOBRE VOCÊS

Formação
Experiências
Mercados
Posicionamento





SOBRE
A DISCIPLINA

RECUPERAÇÃO DE INFORMAÇÃO

75 H

OBJETIVO GERAL

Conhecer os conceitos, fundamentos, aplicações e potencial da RI - Recuperação de Informação e dos Sistemas de Recomendação, com vistas a identificar, avaliar e selecionar adequadamente os modelos de RI mais adequados para a sua efetiva aplicação.



SOBRE
A DISCIPLINA

RECUPERAÇÃO DE INFORMAÇÃO

75 H

OBJETIVO ESPECÍFICOS

- Conhecer o conceito e o funcionamento da Recuperação de Informação (RI) e Sistemas de Recomendação (SR)
- Analisar os impactos que tais tecnologias podem impor sobre mercados;
- Vivenciar, através de bootcamps, momentos de ideação usando tais tecnologias para propor produtos de software em mercados variados;
- Aprender sobre os critérios de escolha a serem considerados durante a seleção de um modelo de RI ou SR.



SOBRE
A DISCIPLINA

RECUPERAÇÃO DE INFORMAÇÃO

75 H

METODOLOGIA DO ENSINO APRENDIZAGEM

As atividades serão baseadas em momentos interativos que privilegiam as relações professor-aluno e aluno-aluno, tais momentos serão construídos utilizando os seguintes procedimentos:

- a) Aulas expositivas-dialogadas;
- b) Apresentação de vídeos;
- c) Uso de ferramentas para experimentação de tecnologias;
- d) Atividades práticas em grupo (Bootcamps)
- e) Leitura de materiais complementares



SOBRE
A DISCIPLINA

RECUPERAÇÃO DE INFORMAÇÃO

75 H

AVALIAÇÕES

- Prova
- Participação nos bootcamps
- Artigo

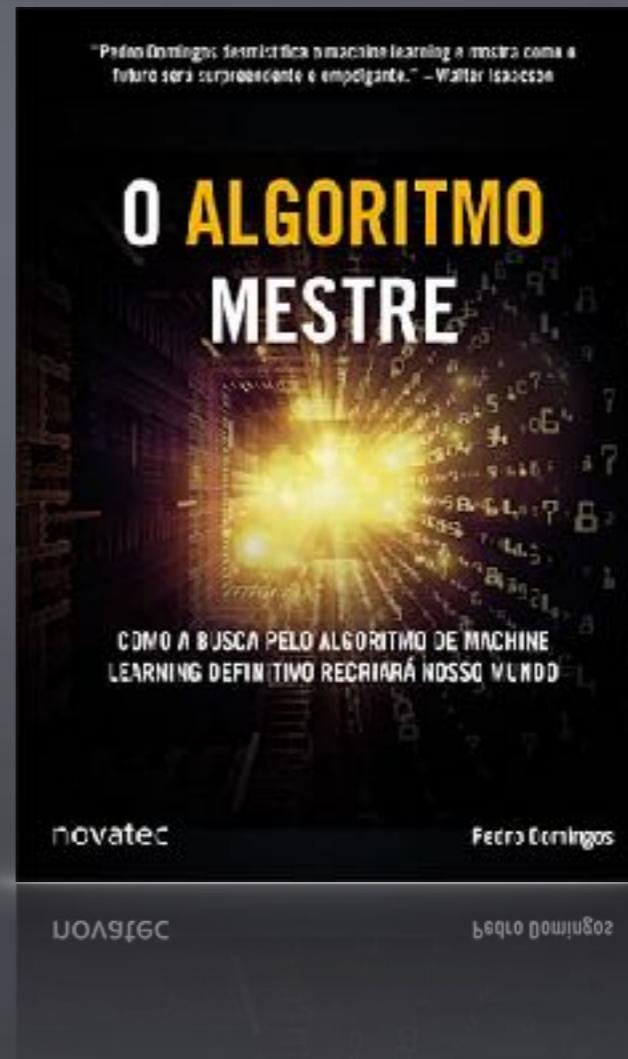
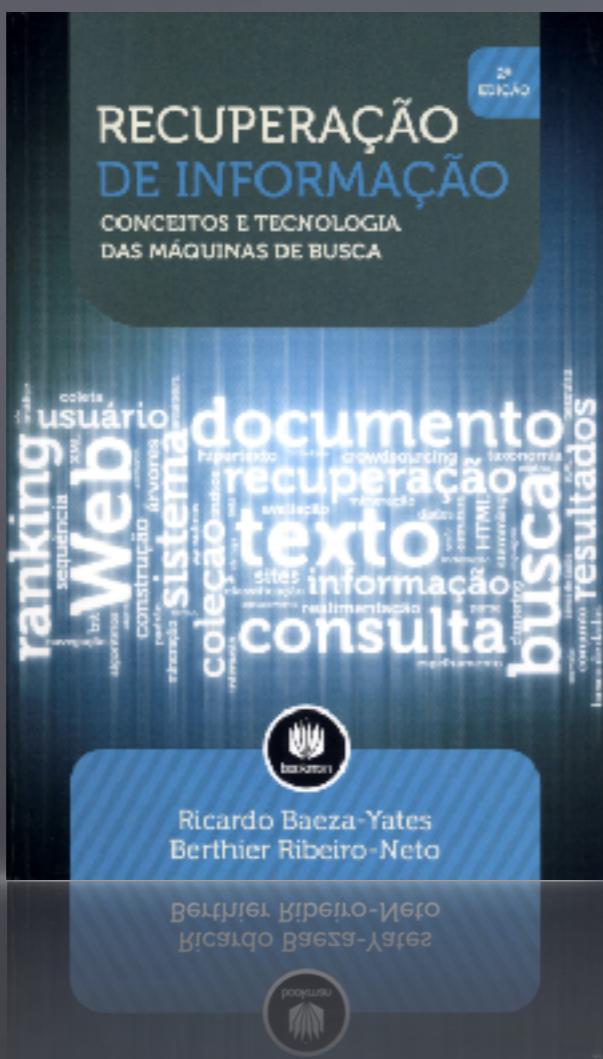


SOBRE
A DISCIPLINA

RECUPERAÇÃO DE INFORMAÇÃO

75 H

BIBLIOGRAFIAS





TECNOLOGIAS EMERGENTES



TECNOLOGIAS EMERGENTES

“Tecnologias capazes de causar grandes impactos sobre o mundo, e sem elas o mundo como conhecemos não existiria”



TECNOLOGIAS
EMERGENTES

QUAIS SÃO
AS TECNOLOGIAS?

- Big Data
- Machine e Deep Learning
- Smart Cities
- Recuperação da Informação e modelos de Recomendação
- Internet of Things
- Visão computacional e reconhecimento de padrões
- Blockchain



A ERA DOS ALGORITMOS



A ERA DOS ALGORITMOS

- Resolvem os cálculos em calculadoras
- Tornam as TV's inteligentes
- Configuram eletrodomésticos
- Permitem aos brinquedos serem interativos
- Pilotam aeronaves
- Gerenciam os sistemas dos carros: Cambio automático, Injeção eletrônica, Controle de tração, Start/Stop e dezenas de outros mecanismos.
- Gerenciam fábricas
- Ganham guerras
- Elegem presidentes

10 anos atrás a maioria da população nunca tinha ouvido falar em algoritmos



A ERA DOS ALGORITMOS

QUAL O VALOR DE UM ALGORITMO?

Um bom algoritmo:

- Usa menos poder computacional (maior eficiência, menos tempo, menos recursos, menos custos)
- Permite uma maior taxa de conversão (mais cliques sobre propagandas)
- Encontra mais documentos semelhantes com a busca do usuário
- Recomenda produtos que o cliente nem sabia que queria comprar



A ERA DOS ALGORITMOS

QUAL O VALOR DE UM ALGORITMO?

Um minúsculo erro em um algoritmo faz:

- Milhões de pessoas ficarem sem energia elétrica
- Um fábrica parar completamente sua produção
- Centenas de vidas perdidas num acidente aéreo
- Um foguete de bilhões de dólares explodir
- Diagnósticos médicos errados
- Semáforos gerenciáveis causarem um colapso em uma cidade
- Um investidor perder dinheiro
- O mercado de ações desabar



A ERA DOS
ALGORITMOS

QUAL O VALOR
DE UM ALGORITMO?

Qual o ativo mais valioso da Google?

- Bens imóveis
- Marca
- Equity (cotas) de outras empresas
- Saldo bancário



A ERA DOS ALGORITMOS

QUAL O VALOR DE UM ALGORITMO?

Um pouco de história

- Prof. Dr. Berthier Ribeiro-Neto - UFMG
- Spin-off Educacional - Akwan
- A Google fez a aquisição da Akwan
- US\$ 225.000.000,00
- 100 doutores em Engenheiros da Computação
- Os demais funcionários não tem acesso à área da engenharia



Google
**COMMUNITY
SUMMIT**

**Mas qual a
resposta para
a pergunta
mesmo ?**





A ERA DOS
ALGORITMOS

QUAL O VALOR
DE UM ALGORITMO?

Qual o ativo mais valioso da
Google?

Algoritmo de Busca



A ERA DOS
ALGORITMOS

QUAL O VALOR
DE UM ALGORITMO?



Números

100 Bilhões de pesquisas mensais

20 Bilhões de sites varridos diariamente

60 Trilhões de endereços (URL)



Números

100 Petabytes de dados armazenados

146 Idiomas



A ERA DOS
ALGORITMOS

QUAL O VALOR
DE UM ALGORITMO?

Indexação por semântica latente Latent semantic Indexing

- Os demais métodos de R.I são baseados em informações explícitas.
- O LSI Encontra uma estrutura semântica entre uma consulta e os termos indexados.
- Traz à tona uma semântica latente (oculta) nos documentos



A ERA DOS
ALGORITMOS

QUAL O VALOR
DE UM ALGORITMO?

Indexação por semântica latente Latent semantic Indexing

- A estrutura semântica está parcialmente obscurecida pela variabilidade da escolha de termos.
carro, veículo, automóvel
- A relevância dos documentos e termos em relação a consulta é estimada considerando a informação extraída da co-ocorrência dos termos na coleção
- 665 Melhorias do sistema de busca (2012)



A ERA DOS
ALGORITMOS

QUAL O VALOR
DE UM ALGORITMO?

- Exemplo em java
- <https://github.com/emmanuelXavier/LSI.git>



I.A EM FOCO

CONCEITO

É a área da Ciência da Computação que estuda o desenvolvimento de máquinas capazes de simular a inteligência humana

OBJETIVO

O objetivo da IA é fazer os computadores se comportarem como humanos.



I.A
EM FOCO

Como os humanos se
comportam ?



I.A
EM FOCO

COMO OS HUMANOS SE COMPORTAM ?

- Ver e reconhecer
- Aprender
- Falar
- Ouvir e entender
- Manusear objetos



I.A
EM FOCO

COMO OS HUMANOS
SE COMPORTAM ?

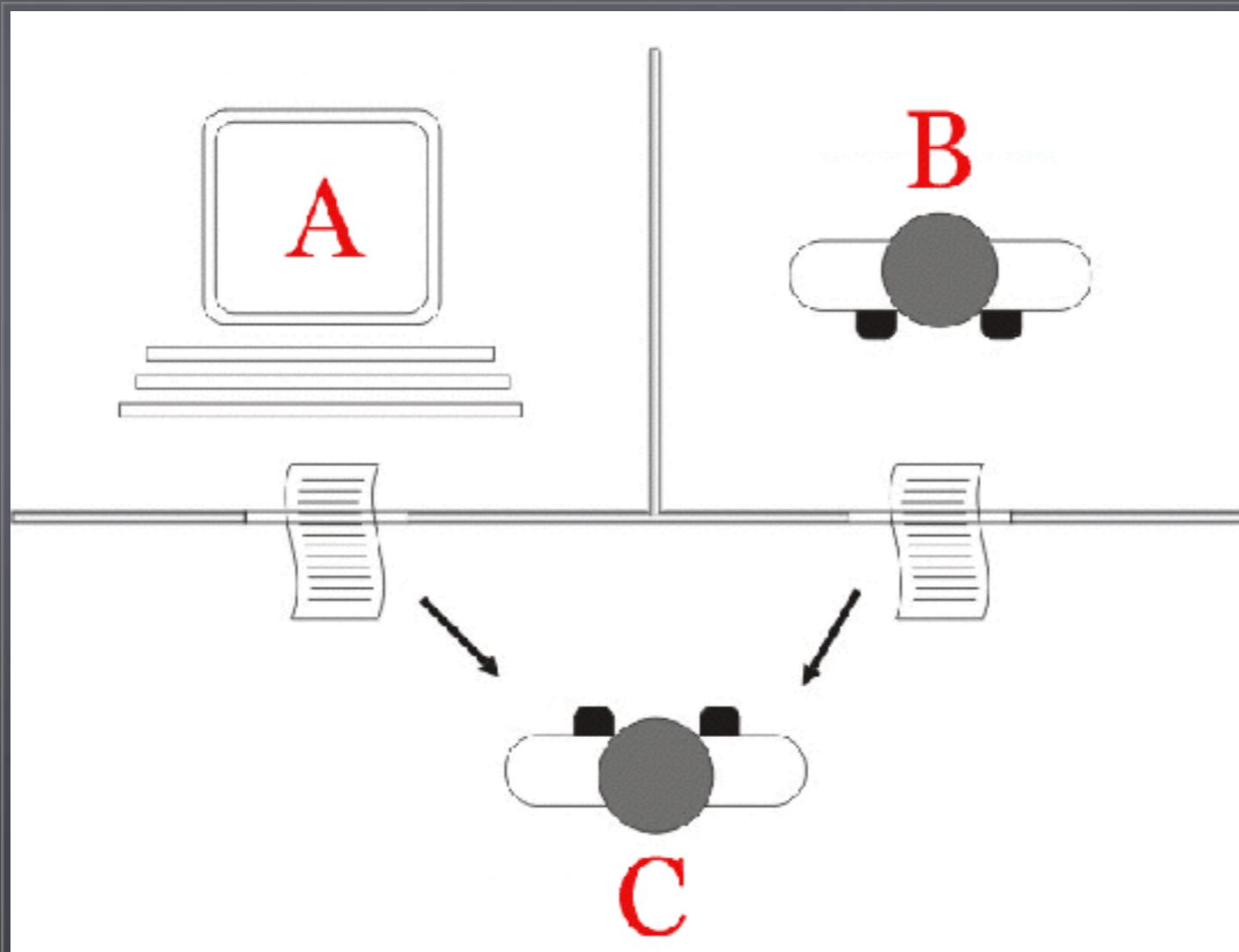
- Ver e reconhecer - Visão computacional e reconhecimento de padrões
- Aprender - Aprendizado de máquina (Machine Learning)
- Falar - Síntese de voz
- Ouvir e entender - Reconhecimento de voz e compreensão de linguagem natural
- Manusear objetos - Robótica



I.A
EM FOCO

COMO MEDIR
A INTELIGÊNCIA ?

Teste de Turing - 1950





I.A
EM FOCO

A I.A ESTÁ PRESENTE
NOSSAS VIDAS ?

- Spam
- Feed de atualizações em redes sociais
- Carros - Injeção de combustível e recirculação de gases de escape
- Tradução de idioma de um situ ou legenda de um vídeo
- Compra de passagem aérea - Bing Travel
- Comandos de voz e assistentes pessoais - Siri, Cortana
- Amazon - Comércio eletrônico



I.A
EM FOCO

A I.A ESTÁ PRESENTE NOSSAS VIDAS ?

- Compras em supermercados - uma I.A. decidiu como mercadorias devem ser distribuídas nas gôndolas e se o freezer de cervejas devem estar próximo do gelo ou das fraldas
- Pagamentos com cartão - Uma I.A. avalia e define se a transação é suspeita.
- Nubank - Uma I.A. avaliou e aprovou sua solicitação do cartão.
- Diminuição de criminalidade - aprendizado estatístico para prever locais com maior probabilidade de ocorrer crimes
- Kinect - Utiliza I.A. para decifrar movimentos
- Netflix



I.A EM FOCO

A I.A ESTÁ PRESENTE NOSSAS VIDAS ?

- Encontrar um par perfeito
- Conhecimento sobre clientes - Eleições - Obama tinha excelentes modelos de eleitores
- Veículos não tripulados e autônomos (mar, ar e em terra)
- Carros autodirigíveis - aprendem a permanecer nas vias sozinhos
- Os algoritmos da NSA descobrem se uma pessoa é um terrorista baseado no que aprendeu sobre os perfis de outros terroristas
- Diagnósticos médicos - aprendeu sobre o par sintomas/diagnósticos
- Correção de digitação - aprendem com usuário



Sistemas Especialistas

Aprendizado de máquina e
compreensão de linguagem natural



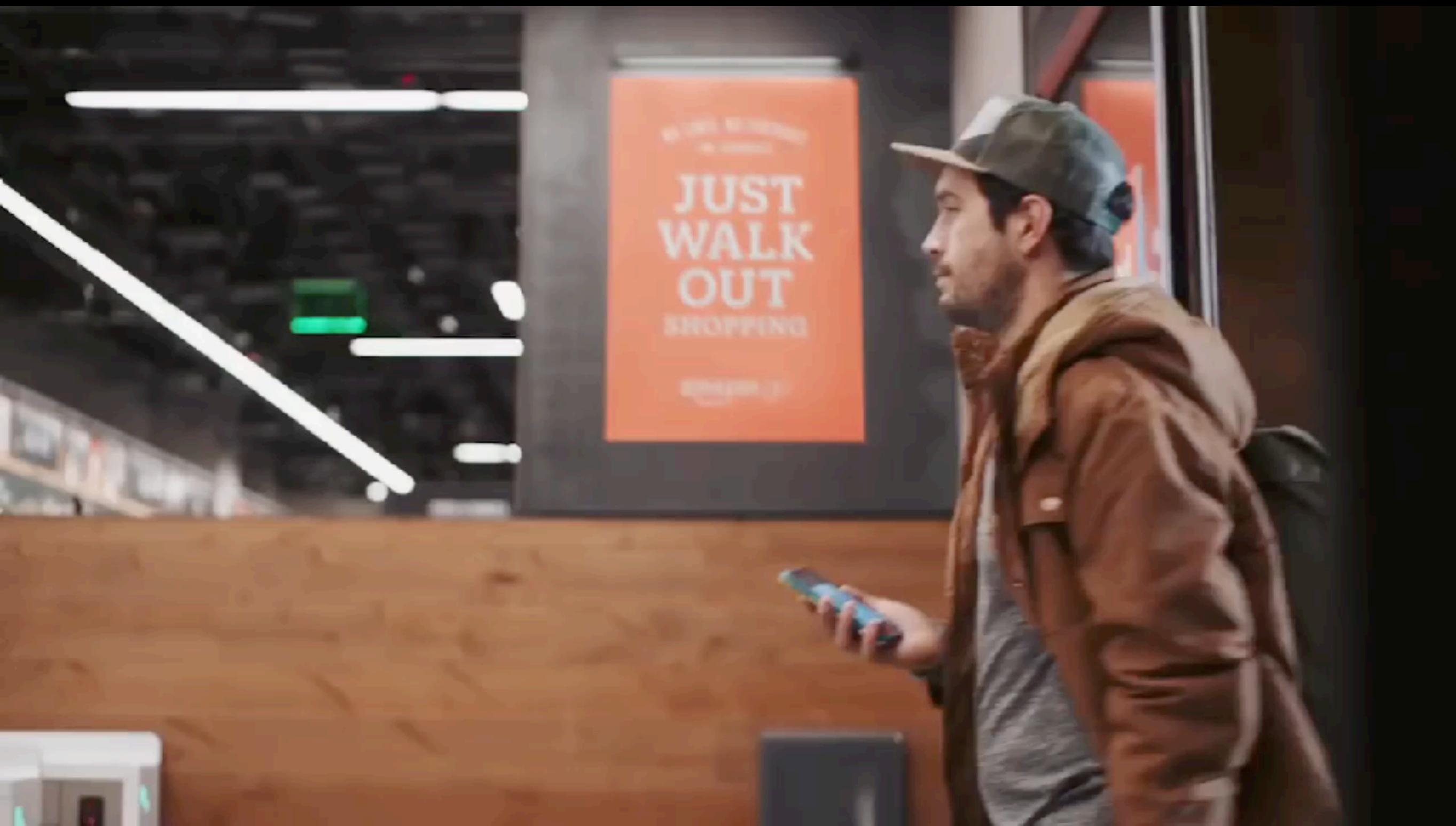
Reconhecimento de padrões

Aprendizado de máquina e
Visão computacional

COZMO®



**Visual Recognition - IBM Watson
Aprendizado de Máquina e visão
computacional**



**Aprendizado de máquina (DEEP LEARNING) e
Visão computacional**

NAO



**Robótica, aprendizado de máquina,
visão computacional, síntese e
reconhecimento de voz**



MACHINE LEARNING

APRENDIZADO DE MÁQUINA

CONCEITO

“O Machine Learning é um subcampo da I.A dedicado ao desenvolvimento de algoritmos e técnicas que permitam ao computador aprender, isto é, aperfeiçoar seu desempenho em alguma tarefa”

(DOMINGO, 2015)

Machine Learning é muito confundido com inteligência artificial, na verdade esse subcampo da I.A. cresceu tanto e foi tão bem-sucedido que ofuscou sua mãe.



MACHINE LEARNING

POR QUE PRECISAMOS DO MACHINE LEARNING?

- **Como os algoritmos funcionam ?**

Todo algoritmo tem uma entrada produz uma saída.

Um conjuntos de dados é dado como entrada, o algoritmo faz o processamento e um resultado é produzido.

- **O Super vilão dos algoritmos - Complexidade**

Complexidade define o custo computacional, ou seja, o uso de recursos computacionais que o algoritmo precisa para fazer sua tarefa

- **Complexidade do espaço:** Quantidade de bits que o algoritmo precisa usar da memória para executar sua tarefa. (Independe da máquina, depende da entrada)

- **Complexidade do tempo:** Quanto tempo o algoritmo leva para executar sua tarefa, ou seja, quantos passos computacionais o algoritmo terá que passar. (Depende da máquina, depende da entrada)



MACHINE LEARNING

POR QUE PRECISAMOS DO MACHINE LEARNING?

- Complexidade do tempo: Classes de crescimento

$O(1)$	- Constante	- Tabelas Hash
$O(\log N)$	- Logarítmica	- Busca binária
$O(N)$	- Linear	- Busca sequencial
$O(N \log N)$	- LogLinear	- Quick Sort
$O(N^2)$	- Quadrático	- Selection Sort
$O(cN)$	- Exponencial	- Algoritmos de força bruta
$O(N!)$	- Fatorial	- Problema do Caixeiro Viajante - A.G



MACHINE LEARNING

POR QUE PRECISAMOS DO MACHINE LEARNING?

- **Complexidade Humana (novo)**

Quando os algoritmos são extremamente complicados para a mente humana resolver.

POR QUE PRECISAMOS DO MACHINE LEARNING?

1º Motivo

Porque existem algoritmos que não conseguimos escrever por serem extremamente complexos.

2º Motivo

Existem aptidões humanas que não conseguimos descrever.

BOOTCAMP

A photograph of two white humanoid robots, possibly NAO models, in an indoor setting. One robot is in the foreground, leaning forward with its right arm extended towards the left. The other robot is in the background, also leaning forward. They are positioned on a light-colored floor against a plain wall.

**Big Data, Recuperação de Informação e
sistemas de recomendação**



BOOTCAMP

BIG DATA, RECUPERAÇÃO DE INFORMAÇÃO E SISTEMAS DE RECOMENDAÇÃO

- Google, Bing, Yahoo, trend.cards, Netflix, Spotify, Amazon, Mercado Livre

Potencial:

- Garçom recomendar pratos
- Descobrir os melhores materiais para estudo
- Fórum de discussão recomendando materiais de estudo
- Recomendar destinos para sistemas de viagens/passagens
- Recomendar roupas baseado no histórico em redes sociais

BOOTCAMP

Big Data, R.I e Sistemas de Recomendação

- **Habilidades avaliadas:**

- **Papel: Cliente Zero**

Participação; capacidade de identificar e expor as dores e ganhos do mercado escolhido.

- **Papel: Negócios**

Participação; capacidade de criar recursos para o produto que possam efetivamente aliviar dores e gerar ganhos para os clientes.

- **Papel: Tecnologia**

Participação; capacidade de avaliar se a tecnologia consegue entregar os recursos oferecidos no produto, verificando sua viabilidade tecnológica



MOTIVAÇÃO

RECUPERAÇÃO DE INFORMAÇÃO

Em 2011 e 2012 foram gerados no mundo 1,7 Zettabytes e 2,7 Zettabytes de dados digitais respectivamente, e a previsão é que em 2015 a massa de dados digitais produzida alcance a casa dos 8 Zettabytes.

(NOGUEIRA et al, 2014)



MOTIVAÇÃO

RECUPERAÇÃO DE INFORMAÇÃO

Durante os anos de 2010 e 2011 gerou-se 90% dos dados disponíveis no mundo até aquele momento.

(NOGUEIRA et al, 2014)



PROBLEMA

RECUPERAÇÃO DE INFORMAÇÃO

Dificuldade e grande consumo de tempo do usuário em encontrar materiais que tenham alto grau de relevância ou utilidade para o seu estudo, nas extensas coleções de documentos disponíveis em bibliotecas físicas ou virtuais, bem como, overload cognitivo resultante do consumo de grande quantidade de informações na busca por conteúdos relevantes.



CONCEITO

RECUPERAÇÃO
DE INFORMAÇÃO

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.

(Berthier, 2013)



CONCEITO

RECUPERAÇÃO DE INFORMAÇÃO

Atualmente a R.I trata da representação, armazenamento, organização e acesso a itens de informação.

- Documentos
- Páginas web
- Catálogos online
- Objetos multimídia
- Outros

Os objetivos iniciais da R.I foram a indexação de textos e busca por documentos úteis em uma coleção.

A área cresceu muito além de seus objetivos iniciais, atualmente a pesquisa em RI inclui modelagem, classificação de textos, arquitetura de sistemas, interfaces de usuário, visualização de dados, filtragem e linguagens



HISTÓRIA

DESENVOLVIMENTOS INICIAIS

Por mais de 5000 anos, a humanidade vem organizando a informação para posterior busca e recuperação.

Desde a biblioteca de Elba, a mais antiga biblioteca conhecida, existente entre 3000 a 2000 a.C, até as enormes bibliotecas digitais que contemplam milhões de títulos, os índices sempre estiveram associados a busca rápida por informação.

Inicialmente criados manualmente os índices continham rótulo, que identificam seus tópicos associados, e por ponteiros, para os documentos que discutem tais tópicos.



HISTÓRIA

DESENVOLVIMENTOS INICIAIS

Os índices são tipicamente projetados e definidos por pesquisadores em biblioteconomia e em ciência da informação.

O surgimento dos computadores modernos possibilitou a construção automática de índices o que acelerou o desenvolvimento da área de RI.

Nos anos 50 aconteceram os desenvolvimentos iniciais na área de RI por meio de esforços de pesquisa de pioneiros como Hans Peter Luhn, Eugene Garfield, Philip Bagley e Calvin Moores (criador do termo Recuperação De Informação)

(Berthier, 2013)



HISTÓRIA

DESENVOLVIMENTOS INICIAIS

Em 1963, Joseph Becker e Robert Hayes publicaram o primeiro livro sobre RI. Vários trabalhos e livros foram publicados após isso.

Em 1978, o primeiro congresso da ACM (Association for Computing Machinery) sobre RI (ACM SIGIR) aconteceu em Rochester, New York.

Em 1979 Van Rijsbergen publicou o livro *Information Retrieval* que apresentava os modelos probabilísticos.
(Berthier, 2013)



HISTÓRIA

DESENVOLVIMENTOS INICIAIS

Em 1983, Salton e McGill publicaram *Introduction to Modern Information Retrieval*, obra clássica de RI que apresentou o modelo vetorial.

Desde então a comunidade de RI cresceu e, atualmente, conta com milhares de professores, pesquisadores, estudantes, engenheiros e profissionais em todo o mundo.

O principal congresso da área, o ACM International Conference on Information Retrieval (ACM SIGIR) recebe centenas de artigos submetidos todo ano.

(Berthier, 2013)



HISTÓRIA

RI EM DESTAKE

Até o início dos anos 90 a RI era vista como uma área de interesse limitada apenas a bibliotecários e a especialistas em informação.



A
WEB

Quem
Quando
Onde
Porque 



HISTÓRIA

RI EM DESTAQUE

A Web, tornou-se um repositório universal da cultura e do conhecimento humano. Milhões de usuários criaram bilhões de documentos que compõem o maior repositório humano do conhecimento na história.

Uma consequência imediata é que encontrar informações úteis na Web não é sempre uma tarefa simples e, normalmente, requer a submissão de uma consulta a uma máquina de busca. Deste modo da noite para o dia a RI ganhou um lugar de destaque junto a outras tecnologias.

(Berthier, 2013)



O PROBLEMA DE RI



RECUPERAÇÃO DE INFORMAÇÃO

O PROBLEMA DE RI

Os usuários de sistemas de RI em geral têm necessidades de informação de diferentes níveis de complexidade.

Necessidade simples

Um link para o site de uma empresa, instituição ou governo

Necessidade complexa

Encontrar todos os documentos que tratam da operação Lava Jato



RECUPERAÇÃO DE INFORMAÇÃO

O PROBLEMA DE RI

"Encontrar todos os documentos que tratam da operação Lava Jato"

Essa descrição não necessariamente fornece a melhor formulação da consulta para um sistema de RI.

O usuário geralmente traduz essa necessidade numa consulta ou em uma sequência de consultas.

A fim de ser efetivo em sua tentativa de satisfazer a necessidade do usuário, o sistema de RI deve de alguma forma “interpretar” o conteúdo dos documentos da coleção, e classificá-los de acordo com o grau de relevância à consulta do usuário.



O OBJETIVO 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

A dificuldade está em saber não apenas como extrair a informação dos documentos, mas também saber como utilizá-la para decidir quanto à sua relevância.



RECUPERAÇÃO DE INFORMAÇÃO

RELEVÂNCIA EM RI

Um ponto importante é que a relevância é um julgamento pessoal que depende da tarefa a ser resolvida e de seu contexto.

1. A relevância pode mudar com o tempo

2. A relevância pode mudar com o local

Nesse sentido nenhum sistema de RI pode fornecer respostas perfeitas a todos os usuários o tempo todo.



CONCEITOS BÁSICOS



CONCEITOS
BÁSICOS

TAREFA DO USUÁRIO
USER TASK

Traduzir sua necessidade de informações em uma consulta, escrita na linguagem fornecida pelo sistema

Implica em especificar um conjunto de palavras que conduzam a semântica de sua necessidade

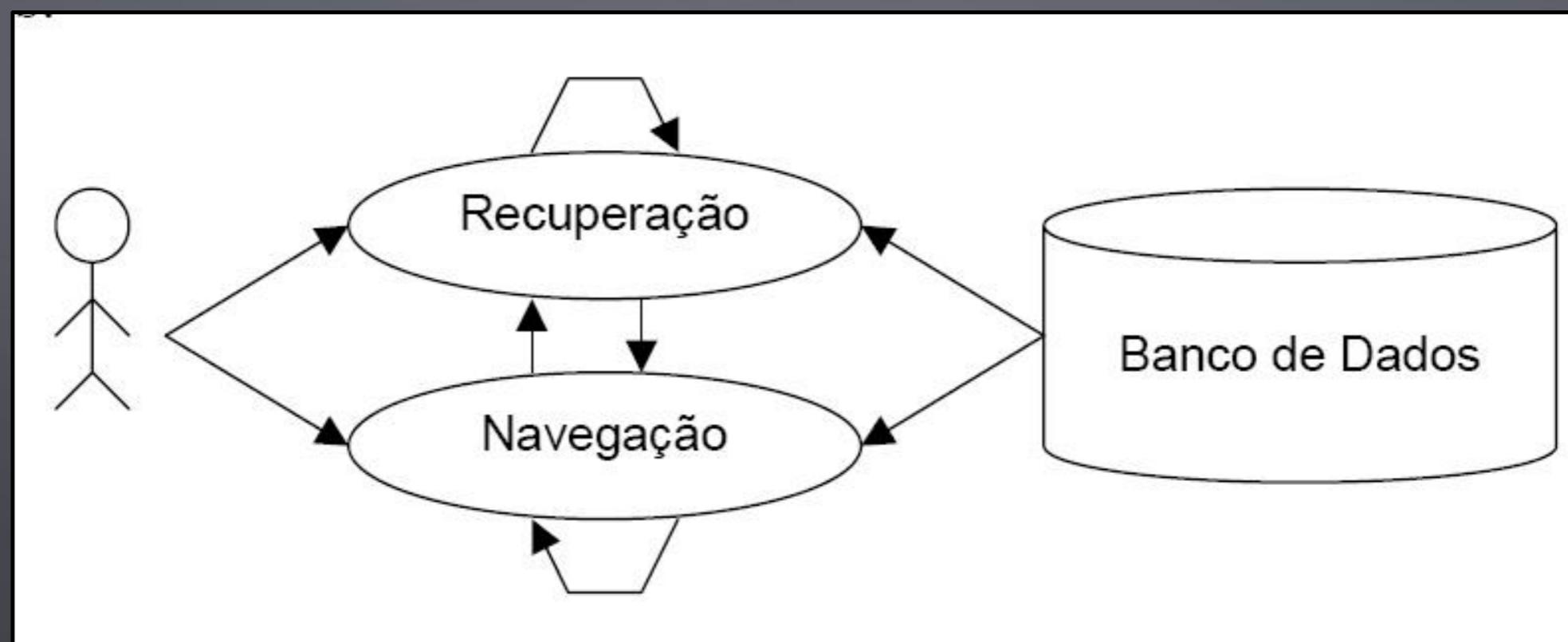
Neste caso, o usuário está buscando por informações úteis executando uma tarefa de recuperação;



CONCEITOS BÁSICOS

TAREFA DO USUÁRIO USER TASK

- Consultas com interesses muito específicos;
- Uma ferramenta que auxilie este usuário a navegar por diversos documentos de uma coleção de documentos é mais interessante, pois será mais abrangente.





CONCEITOS
BÁSICOS

TAREFA DO USUÁRIO
USER TASK

- Sistemas de RI clássicos - recuperação de informação rápida;
- Sistemas de hipertexto - navegação rápida;
- Bibliotecas digitais modernas e interfaces para a web:
 - Devem tentar combinar estas duas tarefas, entretanto esta ainda não é uma abordagem estabelecida;



CONCEITOS
BÁSICOS

VISÃO LÓGICA
DOS DOCUMENTOS

- Documentos em uma coleção são geralmente representados por um conjunto de termos de indexação ou palavras-chaves;
- Podem ser extraídas de duas formas:
 - Automática
 - Selecionados por um especialista humano
- Fornecem uma visão lógica dos Documentos;
- O texto completo seria a especificação mais adequada



CONCEITOS BÁSICOS

VISÃO LÓGICA DOS DOCUMENTOS

- Mesmo os computadores mais modernos, em coleções muito grandes, precisam reduzir o conjunto de palavras-chave representativas. Isto pode ser acompanhado de:
 - Remoção de stopwords: artigos e conectivos
 - Stemming: Substituem palavras flexionadas por seus respectivos radicais)
 - Identificação de substantivos: eliminam adjetivos, advérbios e verbos;
 - Futuramente, também poderá ser empregada a compressão.



MODELOS DE RI



CONCEITOS
BÁSICOS

MODELOS DE RI
INTRODUÇÃO

Os **algoritmos de ranqueamento (classificação)** estão no centro do foco da modelagem de sistemas de recuperação de informações (prevendo quais documentos são relevantes e quais não são).

O objetivo da modelagem em RI é produzir uma função de ranqueamento, ou seja, é produzir uma função que atribui escores a documentos em relação a uma consulta.

(Berthier Ribeiro-Neto)



CONCEITOS
BÁSICOS

MODELOS DE RI
INTRODUÇÃO

A Modelagem em RI preocupa-se com duas tarefas principais:

1. A concepção de um **arcabouço lógico** (formato) 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



CONCEITOS
BÁSICOS

MODELOS DE RI
INTRODUÇÃO

O arcabouço lógico é normalmente baseado em conjuntos, vetores ou em distribuições de probabilidade.

Os sistemas tradicionais de recuperação de informações geralmente adotam **termos de indexação** para indexar e recuperar documentos.

Um **termo de indexação** é uma palavra-chave que possui algum significado próprio (geralmente um substantivo).



CONCEITOS
BÁSICOS

MODELOS DE RI
INTRODUÇÃO

Vantagens

- Simples
- Eficiente
- A semântica dos documentos e da necessidade de informação do usuário pode ser naturalmente expressa através de conjuntos de termos de indexação.



CONCEITOS
BÁSICOS

MODELOS DE RI
INTRODUÇÃO

Desvantagens

- Expressar a intenção da consulta usando poucas palavras restringe a semântica do que pode ser expresso. Assim, não é raro que os documentos recuperados em resposta a consulta do usuário sejam freqüentemente irrelevantes.
- Os usuários não possuem conhecimento sobre como formular suas consultas. A insatisfação de usuários de máquinas de busca com os resultados que obtêm é um sintoma disso.



CONCEITOS
BÁSICOS

MODELOS DE RI
INTRODUÇÃO

A função de ranqueamento é a implementação de um algoritmo preditivo que almeja aproximar-se da opinião de uma grande fração dos usuários quanto à relevância dos resultados recuperados.

É papel também da função de ranqueamento estabelecer um ordenamento dos documentos recuperados.

Documentos que aparecem no topo desse ordenamento são considerados como tendo mais chance de serem relevantes.



CONCEITOS
BÁSICOS

MODELOS DE RI
INTRODUÇÃO

Um algoritmo de ranqueamento opera de acordo com idéia básica (premissa) a respeito da noção de relevância de um documento.

Diferentes conjuntos de premissas (sobre a relevância do documento) produzem modelos de RI distintos.

O Modelo de RI adotado determina as predições do que é e do que não é relevante (noção de relevância).



MODELOS DE RI

CARACTERIZAÇÃO DE UM MODELO

DEFINIÇÃO

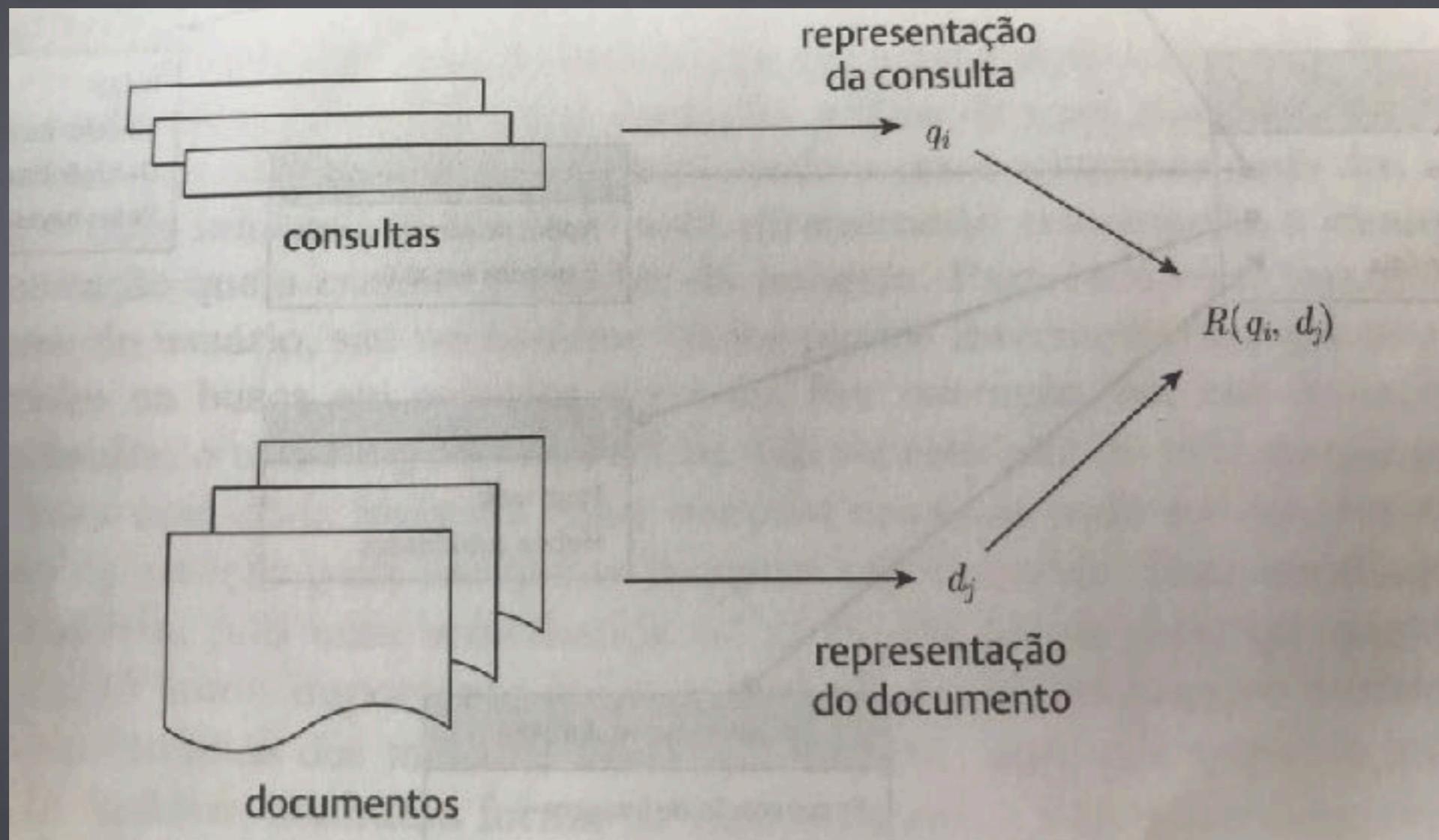
Um modelo de RI é uma quidrupla $[D, Q, F, R(qi, dj)]$ onde:

1. **D** é um conjunto composto por representações (visões lógicas) dos documentos da coleção.
2. **Q** é conjunto composto por representações das necessidades de informação dos usuários (consultas).
3. **F** é um arcabouço lógico para modelar as representações dos documentos
4. **R(qi, dj)** é uma função de ranqueamento que associa um número real à uma representação de uma consulta qi e à representação de um documento dj , assim define um ordenamento entre os documentos em relação à consulta qi .



MODELOS
DE RI

CARACTERIZAÇÃO
DE UM MODELO

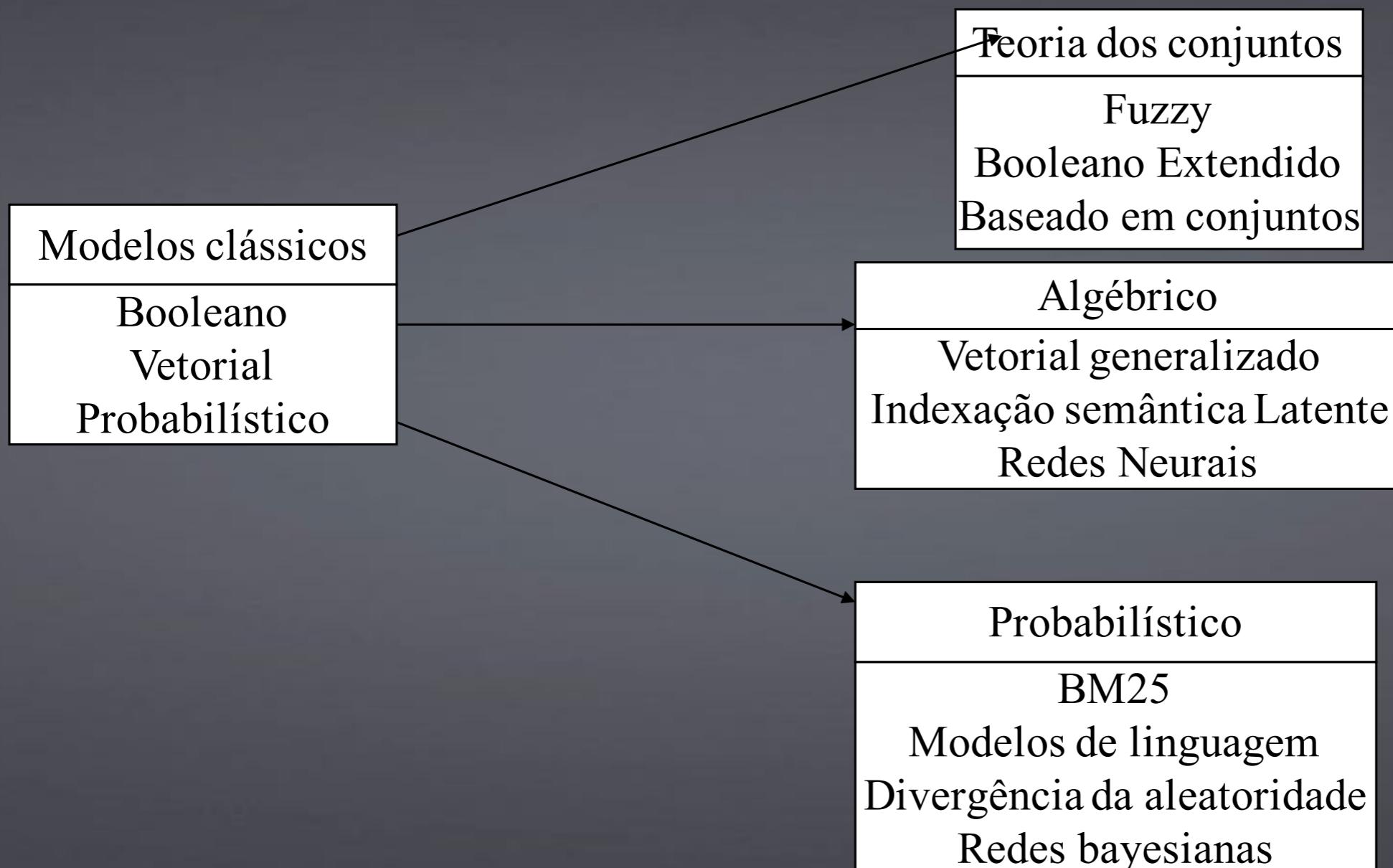


(Berthier Ribeiro-Neto)



MODELOS DE RI

TAXONOMIA DE MODELOS



(Berthier Ribeiro-Neto)



MODELOS
DE RI

TAXONOMIA
DE MODELOS

Os três modelos clássicos de RI (booleano, vetorial e probabilístico) utilizam premissas distintas para definir a relevância dos documentos de uma coleção em relação a uma consulta.

No modelo Booleano, os termos de indexação não possuem peso algum associado, são simplesmente elementos de um conjunto (teoria dos conjuntos).

Nos modelos vetorial e probabilístico, os termos de indexação possuem pesos associados com o objetivo de melhorar o ordenamento dos documentos.



MODELOS
DE RI

TAXONOMIA
DE MODELOS

“

Um termo de indexação é uma palavra-chave que possui algum significado próprio (geralmente um substantivo).”

Para bibliotecários e cientistas da informação, termos de indexação são grupos de palavras pré-selecionadas que representam conceitos-chave (ou tópicos) em um documento.



MODELOS
DE RI

TAXONOMIA
DE MODELOS

Objetivo final na utilização de termos de indexação é descobrir o conteúdo dos documentos. Assim, os termos são principalmente substantivos, uma vez que substantivos possuem significado próprio.

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



MODELOS
DE RI

TAXONOMIA
DE MODELOS

DEFINIÇÃO

VOCABULÁRIO

Considere t como o número de termos de indexação distintos na coleção de documentos, e ki como um termo de indexação qualquer. $V = \{k_1, \dots, k_t\}$ 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 .

(Berthier Ribeiro-Neto)



MODELOS
DE RI

TAXONOMIA
DE MODELOS

O vocabulário é um importante componente da coleção, pois identifica todos os termos de indexação.

À medida que a coleção aumenta, o tamanho do vocabulário aumenta devido a erros de grafia, e termos de outros idiomas.



MODELOS
DE RI

TAXONOMIA
DE MODELOS

MATRIZ DE TERMOS E DOCUMENTOS

A ocorrência de um termo em um documento estabelece uma relação entre eles.

A matriz de termos e documentos representa as relações entre os termos e documentos da coleção.



MODELOS
DE RI

TAXONOMIA
DE MODELOS

MATRIZ DE TERMOS E DOCUMENTOS

	Doc 1	Doc 2	Doc 3	Doc 4
Term 1	1	0	1	0
Term 2	1	1	0	0
Term 3	0	1	1	1
Term 4	1	1	1	1



MODELOS
DE RI

TAXONOMIA
DE MODELOS

MATRIZ DE TERMOS E DOCUMENTOS

As relações entre termos e documentos podem ser quantificadas, por exemplo pela frequência do termo no documento.

$$\begin{matrix} & d_1 & d_2 \\ k_1 & \left[\begin{matrix} f_{1,1} & f_{1,2} \\ f_{2,1} & f_{2,2} \\ f_{3,1} & f_{3,2} \end{matrix} \right] \\ k_2 & \\ k_3 & \end{matrix}$$

(Berthier Ribeiro-Neto)