Universidade Federal de Santa Maria
1960
Frederico Westphalen

Introdução à Recuperação de Informações

Tópicos Especiais em Recuperação de Informações

Profª. Solange Pertile

18/09/15

Objetivos da disciplina

- ✓ Recuperação de Informações
- ✓ Propor soluções
- ✓ Avaliar sistemas de recuperação de informações

Aulas

- ✓ 40% - Teoria
- ✓ 60% - Prática

Avaliação

- ✓ Trabalho 2 - Prático = 3 pontos
- ✓ Trabalho 1= 2 pontos
- ✓ 1 Prova = 5 pontos

Bibliografia da disciplina

Universidade Federal de Santa Maria
1960
Frederico Westphalen

Objetivos desta aula

- ✓ Motivação/Introdução
- ✓ Definição
- ✓ Sistema de Recuperação de Informações ... ??

Objetivos desta aula

RI é recuperação de dados perdidos (deletados)??

Universidade Federal de Santa Maria
1960
Frederico Westphalen

Objetivos desta aula

- ✓ Motivação/Introdução
- ✓ Definição
- ✓ Sistema de Recuperação de Informações Textuais
- ✓ Diferenciar entre Sistemas de Bancos de Dados e RI
- ✓ Introduzir sub-áreas e áreas relacionadas

Motivação

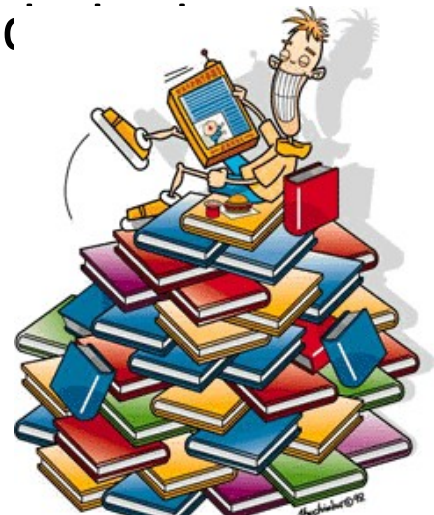
- ✓ O problema: "a tarefa massiva de tornar mais acessível, um acervo crescente de conhecimento". VANNEVAR BUSH (1945)
  - Explosão informacional
  - Importância estratégica da informação

Motivação

- ✓ Documentos digitais de conteúdo processável por computador (desde 1980)
- ✓ Web como repositório digital mundial de informação (desde 1990)
- ✓ Necessidade de condensar e organizar a informação de acordo com necessidades e objetivos para recuperação posterior (OTLET, 1934) .

O que é RI?

- ✓ Recuperação de Informações (RI) trata da representação, armazenamento, organização e acesso a elementos de informação. (Baeza-Yates 2013)
- ✓ Encontrar material de natureza não-estruturada que satisfaz uma informação requerida a partir de grandes coleções. (Mainning 2010)
- ✓A tarefa de um sistema de RI é recuperar documentos (ou textos) com conteúdo que seja relevante à necessi[illegible] informação do usuário (Spark-Jones 1997)



O que é RI?

- ✓ Recuperação de Informações (RI) trata da representação, armazenamento, organização e acesso a elementos de informação. (Baeza-Yates 2013)
- ✓ Encontrar material de natureza não-estruturada que satisfaz uma informação requerida a partir de grandes coleções. (Mainning 2010)
- ✓A tarefa de um sistema de RI é recuperar documentos (ou textos) com conteúdo que seja relevante à necessidade de usuário (Spark-Jones 1997)

Relevante???

Objetivos inicias de RI

- ✓ 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.

Objetivos inicias de RI

- ✓ 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.

Índices???

Universidade Federal de Santa Maria
1960
Frederico Westphalen

Objetivos inicias de RI

- ✓ 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.
  - ✓ **Índices**: coleção de termos que indicam o local onde a informação desejada pode ser localizada.

Universidade Federal de Santa Maria
1960
Frederico Westphalen

Objetivos atuais de RI

- ✓ 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 IEEEXplore

Pesquisa em RI

- ✓A pesquisa em RI concentra-se em capturar a informação fornecida pelo usuário e melhorar o resultado produzido pelo sistema.

  - ✓Centrada no computador
  - ✓Centrada no usuário

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

Centrada no usuário

- ✓Consiste, principalmente, estudar:
  - ✓ O comportamento do usuário
  - ✓ Entender suas principais necessidades
  - ✓ Determinar como esse entendimento afeta a organização e a operação do sistema de recuperação

Necessidade de informação do usuário

- ✓Encontre todas as páginas (documentos) que contenham informações sobre equipes de tênis de universidades que:

  (1) sejam mantidos por uma universidade nos EUA e

  (2) participe do torneios de tênis da NCAA

Para ser relevante, a página deve conter informações sobre o ranking nacional do time nos últimos 3 anos e o email ou número de telefone do treinador do time.

Tarefa típica de um sistema RI

- ✓ **Dados**
  - ✓ Um corpus de documentos e
  - ✓ Uma consulta do usuário
- ✓ **Encontrar**
  - ✓ Um conjunto (ordenados) de documentos que são relevantes para a consulta

Adaptado aulas Prof. Viviane Moreira (UFRGS)

Sistemas de RI

- ✓ Etapas principais na construção do SRI:
  - ✓ Aquisição (seleção) dos documentos
  - ✓ Preparação dos documentos
  - ✓ Indexação dos documentos
  - ✓ Armazenamento
  - ✓ Recuperação
  - ✓ Busca (casamento com a consulta do usuário)
  - ✓ Ordenação dos documentos recuperados

Adaptado aulas Prof. Viviane Moreira (UFRGS)

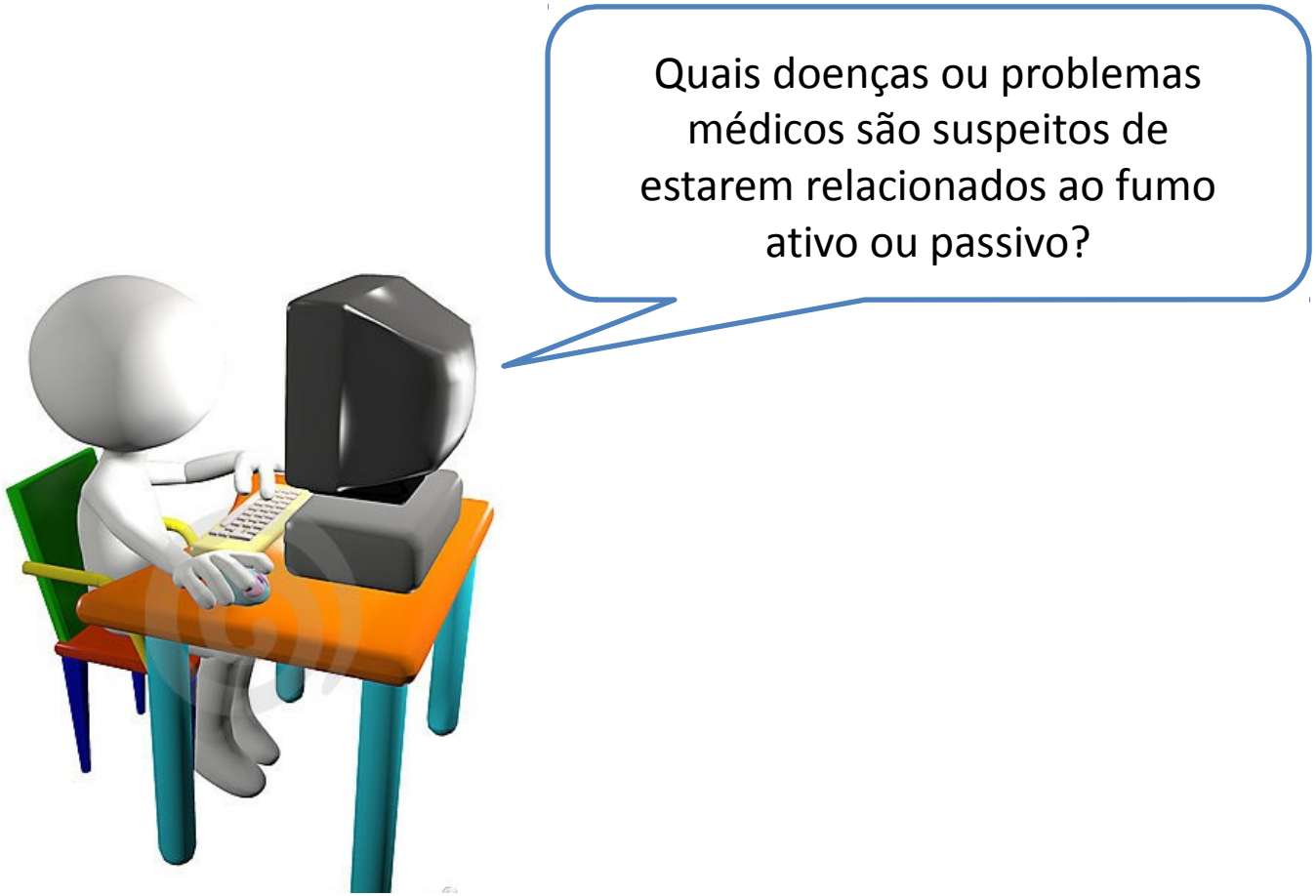
Sistemas RI

"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."

Adaptado aulas Prof. Viviane Moreira (UFRGS)

Processo de RI

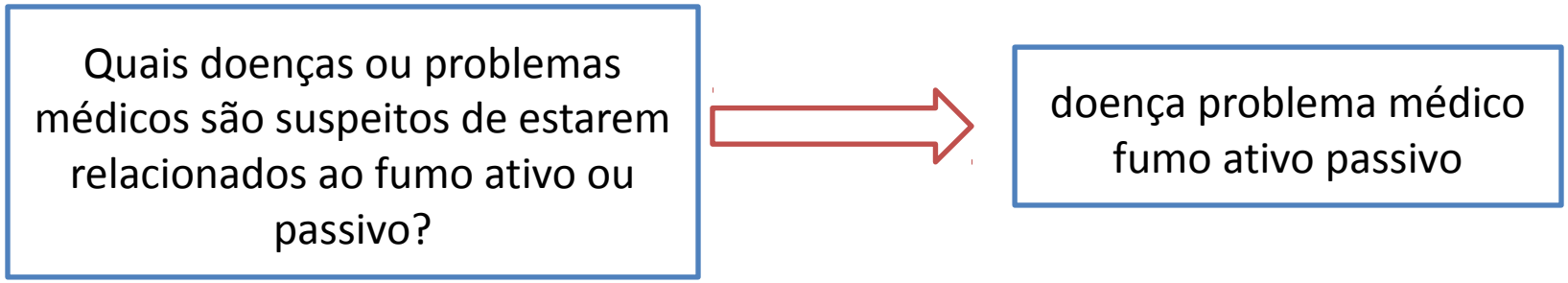
1. O usuário tem uma necessidade de informação



Adaptado aulas Prof. Viviane Moreira (UFRGS)

Processo de RI

2. O usuário tipicamente precisa traduzir sua necessidade de informação em forma de uma consulta (palavras chave).



3. A consulta é submetida a um sistema de RI.

Adaptado aulas Prof. Viviane Moreira (UFRGS)

Processo de RI

4. O sistema de RI processa esta consulta e devolve uma lista de documentos classificada em ordem decrescente de similaridade (parte mais critica de um sistema RI).

97%	**PUCRS :: Hipertexto** O **médico** Aloysio Achutti entende que os **problemas relacionados** ao tabagismo ... O **fumo passivo** é considerado atualmente um grave **problema** de saúde do mundo. ...
92%	**Revista de Saúde Pública - Prevalence of asthma and asthma ...** O tabagismo, **ativo** e **passivo,** tem sido associado com tosse e início tardio ... acesso ao cuidado **médico** e a conseqüente omissão do diagnóstico da **doença** têm ...
90%	**Em tempos que o hábito de fumar vem sendo combatido em muitos ...** câncer de pulmão são atribuídos ao **fumo ativo** e 5% ao **fumo passivo**), ... **doença** tabaco-relacionada, sem nunca ter tragado um cigarro **ou**. fumado um charuto. ...
89%	**Cidade Verde.com :: TV Cidade Verde _ Afiliada SBT em Teresina** Receita sem **médico.** OMS alerta para os **problemas** causados pela ... No Brasil, estima-se entre 80 e 100 mil óbitos **relacionados** ao **fumo** por ano. . .

Adaptado aulas Prof. Viviane Moreira (UFRGS)

Componentes de um sistema RI

Consulta

Processa Texto

Busca nos índices

Recupera documentos

Produz documentos ranqueados

Resultados

Tokenization, remoção de stopwords, stemming e peso dos termos

Arquivo invertido

Coleção de documentos

Adaptado aulas Prof. Viviane Moreira (UFRGS)

Componentes de um sistema RI

Consulta

Processa Texto

Busca nos índices

Recupera documentos

Produz documentos ranqueados

Resultados

- Texto não estruturado
- Dados semi-estruturados
- Imagem
- Som
- Vídeo
- Cadeias de DNA
- ...

Coleção de documentos

Adaptado aulas Prof. Viviane Moreira (UFRGS)

Componentes de um sistema RI

Consulta

Processa Texto

Busca nos índices

Recupera documentos

Produz documentos ranqueados

Resultados

- **Texto não estruturado**
- Dados semi-estruturados
- Imagem
- Som
- Vídeo
- Cadeias de DNA
- ...

Coleção de documentos

Adaptado aulas Prof. Viviane Moreira (UFRGS)

RI – Fácil ou difícil?

Adaptado aulas Prof. Viviane Moreira (UFRGS)

RI – Fácil ou difícil?

- ✓ Consultas ambíguas
- ✓ Como extrair informações dos documentos?
- ✓ 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.
- ✓ Métodos intuitivos nem sempre melhoram os resultados

Adaptado aulas Prof. Viviane Moreira (UFRGS)

RI x Recuperação de dados

Recuperação de Dados (SGBD)

- Tabelas – dados estruturados
- Consultas claras e precisas expressas em uma linguagem (SQL)
- Consultas recuperam todas as tuplas relevantes e todas as tuplas recuperadas são relevantes
- Todas as tuplas recuperadas são igualmente relevantes à consulta

Recuperação de Informações

- Documentos – pouca estrutura
- Consultas vagas e ambíguas expressas por meio de palavras chave
- Consultas não recuperam todos os itens relevantes e nem todos os recuperados são relevantes.
- Documentos apresentados em ordem decrescente de relevância

Adaptado aulas Prof. Viviane Moreira (UFRGS)

Breve Histórico - Nascimento

- ✓Necessidade de organizar a informação
- ✓ Idéias de Vannevar Bush

> *Consider a future device for individual use, which is a sort of mechanized private file and library. It needs a name, and, to coin one at random, "**memex**" will do. A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory.*
>
> Vannevar Bush, As We May Think, The Atlantic Monthly, July 1945

- ✓ 1950 - Termo *Information Retrieval* usado pela primeira vez por Calvin Mooers
- ✓ 1958 - *International Conference in Scientific Information* (Washington) marcou o início da IR

Breve Histórico – Anos 60

- ✓ Desenvolvimento dos primeiros sistemas de informação de grande escala
- ✓ Aparecimento dos sistemas para bibliotecas (*Dialog*)
- ✓ Surgimento da idéia de *"free-text searching"*
- ✓ Propostas de medidas de avaliação (Precisão e Revocação)
- ✓ Criação das primeiras coleções de teste e realização de extensivos experimentos (Cranfield)
- ✓ Criação de novas técnicas de recuperação como *relevance feedback*
- ✓ 1968 – Salton lança seu primeiro livro – proposta do modelo vetorial

Adaptado aulas Prof. Viviane Moreira (UFRGS)

Breve Histórico – Anos 70

- ✓ Interesse em Bancos de Dados e Sistemas de Automação de Escritórios – declínio da pesquisa em RI
- ✓ Possibilidade de se digitar e armazenar texto em um computador (barateamento dos discos)
- ✓ Possibilidade de submeter uma consulta e receber uma resposta imediatamente
- ✓ Popularização de sistemas de bibliotecas
- ✓ Surgimento dos primeiros sistemas de *full-text retrieval*
- ✓ Proposta do Modelo Probabilístico

Adaptado aulas Prof. Viviane Moreira (UFRGS)

Breve Histórico – Anos 80

- ✓ Preço dos discos continuou a baixar
- ✓ Popularização dos sistemas processadores de texto - mais informação digital disponível.
- ✓ Disponibilidade de textos completos
- ✓ Bibliotecas disponibilizaram catálogos para acesso online
- ✓ Distanciamento entre pesquisa e sistemas comerciais

Adaptado aulas Prof. Viviane Moreira (UFRGS)

Breve Histórico – Anos 90

- ✓ Popularização da Internet
  - 1994 – 3 milhões de usuários
  - 1997 – 100 milhões de usuários
- ✓ Surgimentos dos browsers – Mosaic
- ✓ Aparecimento dos motores de busca – Altavista, Yahoo, Lycos, Infoseek, Microsoft, Google.

Adaptado aulas Prof. Viviane Moreira (UFRGS)

Breve Histórico – Presente

- ✓ 2008 – mais de 20 bilhões de páginas web
- ✓ mais de 1 trilhão de URLs distintas coletadas pelo Google[1]
- ✓ 2014 - 3 bilhões de usuários

[1]http://googleblog.blogspot.com.br/2008/07/we-knew-web-was-big.html

Áreas de RI

- ✓ Processamento de Linguagem Natural
- ✓ Inteligência Artificial
- ✓ Probabilidade e Estatística
- ✓ Interação Homem-Computador

Universidade Federal de Santa Maria
1960
Frederico Westphalen

Temas envolvendo RI

- ✓ Recuperação de imagens
- ✓ RI em vídeos
  - ✓Como localizar uma cena X de 5 segundos dentro de um filme de 2 horas?
- ✓ Recuperação de dados geográficos
  - ✓ Adicionar informação georreferenciada ao processo de IR
- ✓ RI sobre dados semi-estruturados
  - ✓ Crescente disponibilidade de documentos XML
  - ✓ Possibilidade de especificar consultas que combinem conteúdo e estrutura

Universidade Federal de Santa Maria
1960
Frederico Westphalen

Temas envolvendo RI

- ✓ Adversarial RI
  - ✓ RI em coleções que foram manipuladas maliciosamente
  - ✓ SpamIndexing – modificar uma página web para que ela seja melhor ranqueada pelos motores de busca
  - ✓ Quais as técnicas usadas por spammers para promover uma página?
  - ✓ Como os motores de busca podem se defender?
- ✓ PageRank
  - ✓ Algoritmo que atribui importância a páginas Web
  - ✓ É a base do motor de busca do Google

Temas envolvendo RI

- ✓Mineração de Opinião
  - ✓ Análise de sentimentos
  - ✓ Determinar se os comentários/opiniões dos consumidores
  - ✓sobre produtos são positivos ou negativos.
- ✓ RI inteligente
  - ✓Retroalimentação (*Feedback* , contexto do usuário, etc.)
- ✓ Detecção de Plágio
  - ✓ Recuperação de documentos candidatos

Exercícios

1. Cite três motivações para pesquisas na área de recuperação de informações (pelo menos uma delas não deve ter sido citada em aula).
2. Qual o principal objetivo de um sistema RI?
3. Defina o processo de RI.
4. Defina o que é um sistema RI.
5. Defina o conceito de relevância em um sistema de RI.
6. Pesquise e descreva o que foi o “memex”.
7. Pesquise e compare como funcionava os primeiros mecanismos de busca em relação aos atuais.

Adaptado aulas Prof. Viviane Moreira (UFRGS)

Trabalho I – Artigo e Apresentação

1. O trabalho é em grupos de 4 alunos
2. Selecionar o tema e informar à professora o quanto antes
   Sugestões de temas estão disponíveis na Aula 1 (o aluno pode sugerir outros)
3. Escrever um artigo sobre o tema de no mínimo 5 páginas e máximo 10
4. Entrega pelo Moodle de uma versão no formato pdf
5. Apresentação de 15-20 min

Adaptado aulas Prof. Viviane Moreira (UFRGS)