Práticas em Banco de Dados

Parte 11 – Recuperação de Informa

Professor Eduardo

Contextualização

- Inicialmente, é importante estabelecer a diferença entre dados estruturados e dados desestruturados (*)
  - Dados **estruturados** são aqueles organizados e representados com uma estrutura rígida, a qual foi previamente planejada para armazená-los, por exemplo um banco de dados.
  - Dados **desestruturados** são aqueles que possuem uma estrutura totalmente inversa dos dados estruturados, sendo flexíveis e dinâmicos ou, até mesmo, sem qualquer estrutura. Um bom exemplo é um documento com textos, imagens, gráficos...

(*) Este conceito será revisto e explorado com mais detalhes em aulas futuras.

Contextualização

- Com o advento da WEB, o volume de dados explodiu.
  - A maior parte desses dados é desestruturada.
  - E o restante está estruturado em uma imensa quantidade de formatos.
- A recuperação de informações lida com problemas de armazenamento, indexação e busca.
  - Estes problemas se agravam com o aumento constante do volume de dados.

Contextualização

- **Recuperação de informações** é o processo recuperar documentos de uma coleção em resposta a uma consulta (ou solicitação de consulta) por um usuário.

- Recuperação de Informações está ligada ao campo de Ciência da Informação não está obrigatoriamente atrelada ao uso de soluções computacionais.
  - Exemplo: antigos sistemas manuais utilizando fichas de catálogo em bibliotecas já aplicavam técnicas de RI (como o uso de palavras-chaves).

Conceito

- **Recuperação de Informações (RI)** é a disciplina que trata da estrutura, análise, organização, armazenamento, pesquisa e recuperação de informações.
- Ela também se aplica ao contexto de dados desestruturados para atender as necessidades de informação dos usuários.
- Um sistema de RI pode ser classificado em diferentes níveis:
  - Por tipos de usuários
  - Por tipos de dados
  - Por tipos de necessidades de informações

Tipos de RI

- **Tipos de Usuários**
  - Usuário **leigo**
    - Tem apenas uma necessidade genérica de informação
    - Não é capaz de elaborar formas relevantes de busca.
    - Exemplo: pesquisadores buscando informações sobre determinado tema; pessoas escolhendo roupas.
  - Usuário **especialista**
    - Tem a visão clara da informação que busca e conhece as especificações exatas da informação
    - É capaz de elaborar formas relevantes de busca para realizar sua tarefa de pesquisa.
    - Exemplo: um curador de acervo de documentos ou um bibliotecário.

Tipos de RI

- **Tipos de Dados**
  - Os sistemas de RI podem ser ajustados a tipos de dados específicos, o que facilita pesquisas direcionadas a determinados temas.
- **Tipos de Informações Necessárias**
  - Pesquisas **Navegacionais**
    - Encontrar um pedaço de informação específica que o usuário necessita.
  - Pesquisas **Informativas**
    - Encontrar atualizações sobre uma informação específica sobre um tópico de interesse do usuário.
  - Pesquisas **Transacionais**
    - Encontrar um site onde acontecem interações que manipulam informações.

Bancos de Dados e Sistemas de RI

BANCOS DE DADOS	SISTEMAS DE RI
Dados estruturados	Dados desestruturados
Controlados por esquemas (tabelas, campos, tipos de dados)	Sem esquema fixo (suporta vários modelos de dado
Modelo de consulta estruturada	Modelo de consulta em formas liv
Consultas retornam dados	Solicitações de pesquisas retornam li ponteiros de documentos
Resultados baseados em combinação exata (não há imprecisão)	Resultados baseados em combinação ap e medidas de eficácia (aceita imprec

Modos de Interação em Sistemas de RI

- **Recuperação**
  - Refere-se à extração de informações relevantes de um repositório de documentos por meio de uma consulta de RI.
  - Interação direcionada por um objetivo pré-definido.
- **Navegação**
  - Atividades de um usuário que visita ou navega por documentos semelhantes ou relacionados, com base na avaliação de relevância feita pelo usuário.
  - A necessidade de informação do usuário não é obrigatoriamente pré-definida e pode ser flexível.

- A busca na WEB combina estes dois aspectos e é uma das principais aplicações de recuperação de informações hoje em dia.

Pesquisa e Análise na WEB

- Na Web, mecanismos de buscas mantém repositórios indexados de p Web e retornam determinadas páginas como respostas a pesquisas feitas usuários. A resposta é baseada em ordem decrescente de pontuaçõ relevância de cada página para uma determinada pesquisa.
- Existem duas abordagens para pesquisa em um sistema de RI:
  - Abordagem **Estatística**
    - Documentos são desmembrados em trechos de texto e cada tre avaliado por sua relevância em relação ao critério de pe fornecido. A avaliação usa técnicas estatísticas para def combinação entre resultados e critérios.
  - Abordagem **Semântica**
    - Recupera informações a partir de uma estratégia baseac conhecimento do tema pesquisado (ainda utiliza métodos estatí mas o foco é no entendimento do conhecimento)

Modelos de Recuperação

- **Modelo Booleano**
  - Documentos são representados como um conjunto de termos e as consultas são formuladas a partir da combinação de termos de pesquisa e operadores booleanos (NOT, AND e OR).
  - É uma abordagem **estatística** com **resultados binários** (verdadeiro ou falso / encontrou ou não encontrou)
  - Não há algoritmos de pontuação de relevância sofisticados. Todos os resultados recuperados possuem importância igual.

Modelos de Recuperação

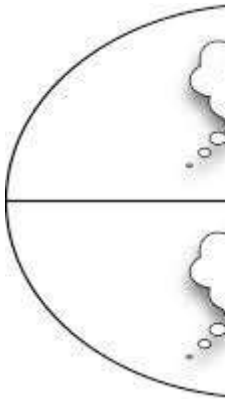
- **Modelo Espaço de Vetor**
  - Neste modelo os termos possuem pesos diferentes, o que permite uma avaliação de relevância e feedbacks mais sofisticados.
  - Documentos são representados como recursos e pesos de recursos de cada termo, armazenados em vetores n-dimensionais de termos.
  - Cada consulta também é especificada como um vetor de termos que é comparado aos vetores de recursos para avaliação de similaridades/relevâncias.
  - É uma abordagem **estatística**.

Modelos de Recuperação

- **Modelo Probabilístico**
  - A solução de espaço de vetor nem sempre oferece respostas com similaridades precisas. Para corrigir isso, o modelo probabilístico pontua os documentos com **probabilidades estimadas de relevância** em relação a consulta realizada.
  - Para decidir essa relevância, o modelo assume que existem **dois conjuntos pré-definidos**: um com resultados relevantes para consulta e outro com resultados não relevantes. A partir disso, calcula as probabilidades de cada documento pertencer a um dos dois conjuntos.
  - Também é uma abordagem **estatística**.

Modelos de Recuperação

- **Modelo Semântico**
  - As abordagens estatísticas, por mais precisas que sejam, sempre oferecem o risco de deixar algum resultado relevante de fora de uma pesquisa, pois não há interpretação de significados.
  - Os modelos semânticos avaliam o nível de coincidência de conceitos e significados em lugar da combinação exata de palavras-chaves.
  - Para isso, utilizam:
    - **Análise morfológica** (verbos, substantivos, adjetivos e outros termos de palavras do discurso)
    - **Análise sintática** (frases do discurso)
    - **Análise semântica** (ambiguidades e sinônimos em palavras, parágrafos, páginas ou documentos inteiros)



Tipos de Consultas em RI

- **Consultas por palavra-chave**
  - Pesquisa a ocorrência de uma ou mais palavras-chave fornecidas usuário.
  - Alguns sistemas removem palavras de ligação que ocorrem com frequência (um/uma, a/o, de/da/do, etc.).
  - A maioria dos sistemas de RI não leva em consideração a ordem co as palavras-chave são fornecidas na consulta.
  - Todos os modelos de recuperação oferecem suporte a consulta palavra-chave.

Tipos de Consultas em RI

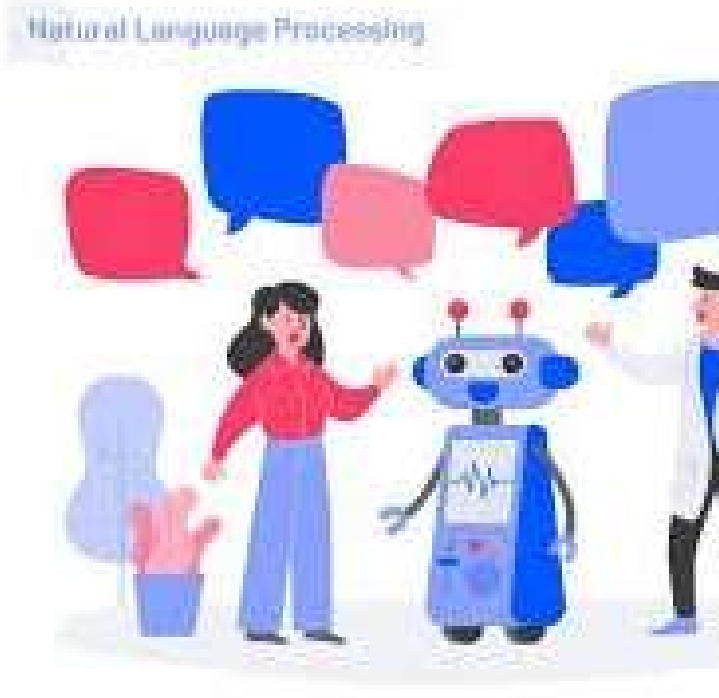
- **Consultas booleanas**
  - Alguns sistemas de RI permitem o uso de operadores booleanos AND, OR, (), + e – em combinação com palavras-chave, permiti elaboração de sentenças lógicas mais complexas.
  - Não é um sistema baseado em pontuação, já que os únicos resu possíveis para um documento é verdadeiro (atende as condiçõ pesquisa) ou falso (não atende).
- **Consultas de frase**
  - Para recuperar documentos utilizando uma frase exata, é preciso frases codificadas em índice ou implementar alguma solução que cata as posições relativas de palavras nos documentos.
  - Cada documento recuperado deve ter pelo menos uma ocorrência da exata em seu conteúdo.

Tipos de Consultas em RI

- **Consultas por proximidade**
  - A consulta considera a proximidade que um dos termos da consulta estar de outro para ser considerado aceitável..
  - Algumas soluções permitem uso de operadores que refinam a pesqu proximidade, como NEAR (perto), ADJ (adjacente) ou AFTER (de porém, quanto melhor o suporte a esses operadores, mais dispendio termos computacionais a pesquisa se torna, por isso é mais indicada documentos pequenos.
- **Consultas por coringas**
  - A pesquisa dá suporte a expressões regulares (wildcards) e padrõ textos.
  - Também tem alto custo computacional, por isso é pouco adotad consulta na Web.

Tipos de Consultas em RI

- **Consultas em linguagem natural**
  - A consulta tenta interpretar textos pesquisados dentro de padrões linguísticos já estabelecidos usando modelos semânticos.
  - É uma área que chama muita atenção e que conta com avanços trazidos pela pesquisa na área de inteligência artificial.



Tendências

- **Busca facetada:** muito usada em sites de e-commerce, é uma experiência integrada de busca e navegação, permite que o usuário explore enquanto refina os dados.
- **Busca social:** cruza experiências de buscas de diversos usuários para melhorar a experiência individual.
- **Busca conversacional:** participantes se engajam em uma conversação que permite uma busca social auxiliada por agentes inteligentes de software, que como resultado de sua aprendizagem, oferecem resultados mais rápidos e mais relevantes aos usuários.

Referências e Links Interessantes

- ELMASRI, Ramez; Navathe, Shamkant B. **Sistema de Banco de Dados**. 7a edição. Pearson. [Recurso eletrônico, Biblioteca Virtual]. Capítulo 27.
- A História da Biblioteca de Alexandria (apresentada por Carl Sagan)
  - https://www.youtube.com/watch?v=TjnE1gV42Jw
- Introdução a RI
  - https://www.youtube.com/watch?v=dGGo-oW5rjA