

UNIVERSIDADE DE BRASÍLIA
Faculdade do Gama

Sistemas de Banco de Dados 2

Tecnologias de Banco de Dados (TI-BD)

Banco de Dados Textual

Max Henrique Barbosa 16/0047013

Brasília, DF

2019

1. Definição

Bancos de dados textual é uma compilação de documentos ou outras informações na forma de um banco de dados no qual o texto de cada documento referenciado está disponível para visualização, impressão ou download online. Além dos documentos de texto, as imagens são frequentemente incluídas, como gráficos, mapas, fotos e diagramas. Um banco de dados de texto pode ser pesquisado por palavra-chave, frase ou ambos.

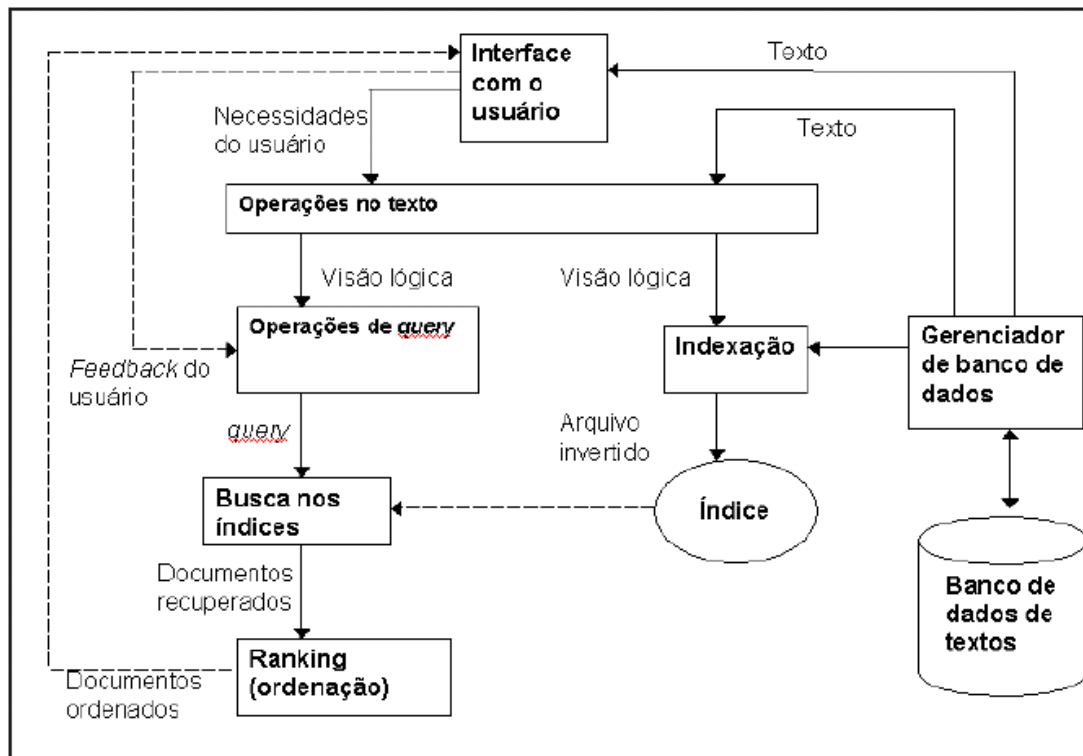
Por se tratar de um banco, que foi criado pela necessidade de se guardar grandes textos como informação, existe a necessidade de serem atrelados a um sistema gerenciador de bancos já que se comportam de forma semelhante aos bancos convencionais. Um sistema de gerência de banco de dados é um software que incorpora as funções de definição, recuperação e alteração que se pode fazer em um banco.

No momento que ocorre uma consulta no banco textual, cada documento deverá ser comparado aos termos da mesma para saber se ele é uma possível resposta. Para que consultas nesse processo sejam mais eficientes é utilizado a estrutura de índice. No geral, um índice é qualquer palavra que apareça no texto. Por conta disso a recuperação de informação está relacionada ao conjunto de termos do texto que será utilizado para a composição do índice. Pode-se definir palavras-chave de forma manual ou automática e utilizar estas como o índice, ou então indexar automaticamente todas as palavras do texto. Geralmente a seleção automática de palavras não consegue obter sucesso na indexação de todas as palavras, por conta disso obtém mais sucesso quando as indexações são realizadas a partir de um padrão manual.

A ideia principal da utilização é a recuperação de informações em meio a uma grande quantidade de textos. Antes de se iniciar o processo de recuperação é necessário submeter cada texto da coleção de dados em um conjunto de operações para que seja gerada sua visão lógica. Após esse processo o sistema gerenciador do banco poderá construir os índices do texto.

Com todos os índices construídos pode-se iniciar o processo de recuperação da informação. Para que isso aconteça o usuário deve descrever

o que deseja obter da coleção de dados, essa mesma descrição também será submetida a um outro conjunto de operações textuais para que seja criada a consulta. A consulta também poderá ser reformulada pelo sistema visando o melhor resultado possível ou pode ser enviada para o próximo passo. Com os termos da consulta obtidos, é então realizada a busca por arquivos e logo após são classificados, através de um ranking, de acordo com a sua relevância.



2. Objetivos principais

Por se tratar de um banco onde seus principais dados são livros, dissertações, periódicos, jornais e outros tipos de documentos de texto seus objetivos principais consiste em armazenar esses textos e devolver, através de um esquema de consulta e relevância, a informação ou o texto mais adequado.

3. Vantagens

As vantagens dos bancos textuais em relação aos bancos relacionais referentes aos mecanismos de pesquisa.

Otimizado para lidar com dados textuais

Os mecanismos de banco de texto são especificamente otimizados para lidar com dados de texto. Em particular, nomes próprios de clientes, cidades, nomes de ruas e outros marcadores geográficos são de natureza textual. Esse tipo de dados textuais geralmente está sujeito a várias grafias válidas e inválidas. Além disso, nomes próprios são frequentemente representados de várias formas não normalizadas.

Estrutura granular do índice

A estrutura de índice dos mecanismos de banco de texto é mais granular, permitindo acesso rápido indexado a palavras e frases específicas.

Operadores de consulta mais flexíveis

Os mecanismos de pesquisa oferecem muito mais operadores de consulta, como idioma, dicionário de sinônimos, correspondências curinga rápidas, similaridade estatística e densidades de palavras, proximidade, Soundex e outras técnicas de correspondência "difusa".

Suporte de pesquisa híbrida

Os mecanismos de pesquisa permitem pesquisas híbridas que podem pesquisar dados em campo tradicionais e dados textuais na mesma consulta.

Ponderação de relevância

Os mecanismos de pesquisa fornecem pesos relevantes para correspondências prováveis, permitindo um ajuste muito mais preciso dos resultados da pesquisa.

4. Diferenças técnicas

Embora existam semelhanças entre o banco de texto e tecnologias relacionais, também existem várias diferenças devido às diferenças fundamentais entre os tipos de dados que estão sendo indexados e a flexibilidade das opções de recuperação. Embora as diferenças possam

apresentar alguns desafios, elas também apresentam a oportunidade de aproveitar os principais recursos da pesquisa de banco de texto para fornecer uma solução inovadora para o problema em questão.

Diferenças no Vocabulário Técnico

Como mencionado, os sistemas relacionais e de banco de texto usam muitos termos semelhantes. Por exemplo, os sistemas relacionais e de banco de texto usam o termo "índice" quase da mesma maneira, mas existem algumas diferenças. Essas diferenças de vocabulário são importantes para pessoas com formação em relacionamento que começam a trabalhar com mecanismos de banco de texto.

Conforme sugerido anteriormente, um usuário ou programa gera uma consulta que é passada para um mecanismo relacional ou de banco de texto, que processa a consulta e retorna um conjunto de resultados. Um sistema relacional retorna linhas de dados em campo. Um sistema de texto retorna metadados que representam documentos.

Um banco de dados ou tabela no mundo relacional é semelhante, mas não idêntico, a uma coleção no mundo do banco de texto. Em alguns casos, uma coleção é mais análoga a uma visão relacional. O que um sistema relacional chama de registro do conjunto de resultados geralmente é chamado de documento por mecanismos de banco de texto.

Outros termos são específicos para mecanismos e fornecedores de banco de texto e não têm significado no ambiente relacional.

Do mesmo modo, alguns conceitos do mundo relacional estão visivelmente ausentes do ambiente de banco de texto. Como discutiremos, os mecanismos de banco de texto não usam junções externas e externas. Além disso, como os mecanismos de banco de texto tendem a servir como um meio somente leitura, há muito menos ênfase nas transações e mecanismos relacionados. As pessoas familiarizadas com os sistemas relacionais podem ficar inicialmente intrigados com a sua ausência, dada a sua importância no mundo dos bancos de dados.

Estrutura de dados

Os mecanismos relacionais geralmente armazenam dados estruturados, de modo que os dados associados sejam armazenados no mesmo registro ou linha, com os componentes organizados em campos ou colunas de informações identificados. Documentos e imagens formatados são mantidos em campos longos especiais, como objetos longos binários, embora o escopo dos comandos que podem ser usados nesses campos longos seja limitado.

Os mecanismos de banco de texto são otimizados para processar documentos formatados e não formatados, mas também devem ser capazes de processar dados estruturados limitados, como títulos, autores e descrições de documentos. Os conjuntos de comandos da maioria dos mecanismos de banco de texto são otimizados para recuperação flexível de documentos, com um conjunto de comandos menos rico em recursos para processar e recuperar os dados da estrutura.

Sintaxe da consulta

Existe uma variedade muito maior de sintaxe da linguagem de consulta nos mecanismos de banco de texto. Muitos mecanismos suportam os operadores clássicos "AND", "OR" e "NOT" de alguma forma; alguns mecanismos até permitem consultas aninhadas complexas usando parênteses ou outra sintaxe de aninhamento. Um pseudo-padrão moderno é a "sintaxe da Internet", que permite que "+" e "-" sejam usados como atalhos para operadores booleanos. Muitos mecanismos de pesquisa realmente oferecem suporte a várias sintaxes que geralmente podem ser configuradas quando uma pesquisa é realizada.

Documentos Virtuais

No coração dos sistemas de banco de texto está o conceito de uma "unidade de recuperação". Geralmente, é um arquivo ou documento, mas quando os dados de origem são de natureza relacional, a unidade de recuperação geralmente é um "documento virtual" composto de dados de várias colunas e tabelas.

Chaves do documento

Como nos sistemas relacionais, documentos de banco de texto são identificados usando um campo-chave. Nos sistemas de banco de texto, a chave do documento pode ser um nome de arquivo ou documento ou um URL. Se o sistema de banco de texto for usado para indexar e pesquisar dados relacionais, a chave do documento geralmente é a chave primária na tabela ou exibição relacional.

5. Desvantagens

Quanto as desvantagens o banco de dados textual é mais vantajoso para a recuperação de dados do que para a pesquisa. Além disso uma desvantagem são as consultas de textos podem retornar textos que não eram relevantes na busca.

Outro ponto negativo é que esses bancos tendem a ter uma cobertura de dados mais limitada. E infelizmente por não ter uma capacidade muito grande para armazenamento não se pode confiar apenas nos dados devolvidos de uma fonte específica.

Comparando a questão das vantagens e desvantagens com os Bancos Relacionais, pode-se dizer que os Bancos Textuais são bastante parecidos, entretanto os dados que costumam guardar são “maiores” em tamanho mas não em quantidade, por conta disso os bancos textuais fazem a separação das informações que retorna por relevância onde bancos relacionais retornariam se a consulta fosse bem detalhada.

6. Exemplos de Uso e Tecnologia

Os bancos de dados de texto são usados pelas bibliotecas de faculdades e universidades como uma conveniência para seus alunos e funcionários. Os bancos de dados em texto são ideais para cursos on-line onde ocorre o armazenamento de apostilas, livros e questões atrelados a um assunto específico onde o aluno permanece em casa e obtém materiais do curso baixando-os da Internet. O acesso a esses bancos de dados

normalmente é restrito ao pessoal registrado ou às pessoas que pagam uma taxa especificada por item visualizado. Os bancos de dados em banco de texto também são usados por algumas empresas, escritórios de advocacia e agências governamentais.

Por ser bastante parecido com base de dados relacionais, os bancos textuais consistem basicamente na implementação e priorização de estruturas de indexação, podendo ser implementado por exemplo em SQL Server ou no MySQL

7. Bibliografia

[Heuser, Carlos Alberto. Projeto de Banco de Dados](#)

Full-text Database

Disponível em: <<https://searchsqlserver.techtarget.com/definition/full-text-database>>. Acesso em: 07 nov. 2019.

Garcia, Marcela Ortega, Recuperação de informações em Bancos de Dados Textuais. 2009. 30f. Instituto de Matemática e Estatística da Universidade de São Paulo

Wiki Full-text Database

Disponível em: <https://en.m.wikipedia.org/wiki/Full-text_database>. Acesso em: 07 nov. 2019.

Understanding Full-text Database

Disponível em: <<https://www.red-gate.com/simple-talk/sql/learn-sql-server/understanding-full-text-indexing-in-sql-server/>>. Acesso em: 07 nov. 2019.

Database Full-text search

Disponível em: <<http://www.ideaeng.com/database-full-text-search-0201>>. Acesso em: 07 nov. 2019.