UNIVERSIDADE DE BRASÍLIA
Faculdade do Gama

Sistemas de Banco de Dados 2

Tecnologias de Banco de Dados (TI-BD)

Banco de Dados Textuais

Kevin Luis Apolinario Batista - 18/0042386

Brasília, DF
2023

A) Introdução

Este trabalho terá como objetivo abordar os principais conceitos relacionados aos bancos de dados textuais. Será apresentada uma definição sobre essa tecnologia e como ela é utilizada para armazenar e gerenciar grandes volumes de dados em formato de texto. Além disso, serão discutidos os objetivos principais dos bancos de dados textuais. Também serão exploradas as vantagens e desvantagens dessa tecnologia, como a capacidade de buscar informações de forma rápida e a complexidade do processo de armazenamento. Por fim, serão apresentados exemplos de uso de bancos de dados textuais em diversos contextos, como na pesquisa acadêmica e no mercado corporativo.

B) Definição da tecnologia

Um texto é uma sequência de símbolos retirados de um alfabeto, e representa uma grande porção das informações disponíveis no mundo eletrônico. Exemplos incluem texto em linguagem natural, como livros, revistas, jornais, bancos de dados de jurisprudência, informações corporativas e conteúdo na web. A combinação de texto e dados estruturados em dados semi estruturados é cada vez mais comum devido à complexidade crescente das aplicações, e esses dados são frequentemente expressos e manipulados em formatos como XML. (Navarro, 2005)

Para lidar com a enorme quantidade de dados textuais disponíveis e permitir buscas rápidas e precisas, os bancos de dados textuais são utilizados. Esses sistemas mantêm coleções de texto e fornecem acesso rápido e preciso a eles. Com o aumento contínuo do volume de dados textuais disponíveis, torna-se cada vez mais importante aprimorar a tecnologia de bancos de dados textuais para permitir que as pessoas encontrem as informações que precisam de maneira eficiente. (Navarro, 2005)

De acordo com Zobel, Moffat e Sacks-Davis (1992), um banco de dados de texto armazena o texto completo de um documento ou conjunto de documentos, indexando cada palavra ou termo do texto para permitir pesquisas de frases ou conceitos específicos.

Esse tipo de banco de dados permite que os usuários pesquisem e recuperem documentos com base no conteúdo de seu texto. Para que a indexação seja eficiente, é necessário utilizar técnicas avançadas que reduzam o espaço necessário para armazenar os índices e acelerem o processo de busca.

C) Principais objetivos da tecnologia

Um dos principais objetivos da área de banco de dados é armazenar e recuperar dados de maneira eficiente. O crescimento do uso de bancos para dados sem estrutura definida desencadeou a necessidade de técnicas diferenciadas. Dentre esses tipos de dados, encontram-se os textos e um exemplo que ilustra esse cenário é a pesquisa em páginas Web, um conjunto volumoso corriqueiramente consultado.

As tecnologias de banco de dados tradicionais não são adequadas para lidar com bancos de dados de texto. Os bancos de dados relacionais estruturam os dados em registros de comprimento fixo cujos valores são atômicos e são pesquisados por igualdade ou intervalos. Não há uma maneira geral de particionar um texto em registros atômicos, e tratar todo o texto como um valor atômico é de pouco interesse para as operações de pesquisa exigidas pelos aplicativos. O mesmo problema ocorre com os tipos de dados multimídia. Daí a necessidade de tecnologia específica para gerenciar textos. (WITTEN; MOFFAT; BELL, 1999, p. 8)

Embora os sistemas de recuperação de texto completo sejam um tipo de banco de dados muito grande, o último termo é geralmente usado para se referir especificamente a bancos de dados convencionais muito grandes, que formam uma área importante de estudo em si. A recuperação de texto completo e os sistemas de banco de dados também fazem parte do campo mais amplo conhecido como recuperação de informações, que pode ser definido livremente como o estudo de métodos e estruturas usados para representar e acessar informações. (WITTEN; MOFFAT; BELL, 1999, p. 8)

D) Vantagens da Tecnologia Pesquisada

- Uso de técnicas de compressão de dados

Em um banco de dados textual, a compressão de texto geralmente envolve técnicas como a codificação de Huffman ou a codificação de dicionário, que reduzem o tamanho do arquivo de texto original por meio da substituição de sequências de caracteres repetidos ou frequentes por códigos mais curtos. A compressão de texto é uma técnica eficaz para reduzir o tamanho de grandes arquivos de texto e pode ser especialmente útil para armazenar grandes quantidades de dados de texto em um único arquivo (WITTEN; MOFFAT; BELL, 1999, p. 23).

Por outro lado, em um banco de dados relacional, a compressão é geralmente feita a nível de registro ou bloco, ao invés de a nível de caracteres. A compressão de registros pode ser eficaz para reduzir o espaço necessário para armazenar colunas com valores repetidos ou de tipos de dados numéricos, além de melhorar as taxas na transferência de dados, e para obter melhor desempenho em aplicações intensivas sobre os dados (ROSA, 2006).

- Indexação do texto

A indexação é um recurso essencial para melhorar o desempenho de consultas em bancos de dados, permitindo que os dados sejam localizados com mais eficiência. Tanto em bancos de dados textuais quanto em bancos de dados relacionais, a indexação é realizada para acelerar o processo de pesquisa e recuperação de dados.

Em bancos de dados textuais, a indexação é geralmente feita por meio de índices de palavras-chave, que permitem que os usuários localizem palavras ou frases específicas no texto. Os índices de palavra-chave são criados a partir de uma lista de palavras distintas no texto e seus locais, permitindo que os usuários realizem pesquisas rápidas e precisas (WITTEN; MOFFAT; BELL, 1999, p. 14).

Já em bancos de dados relacionais, a indexação é geralmente realizada em colunas específicas das tabelas. Os índices em bancos de dados relacionais são usados para acelerar a pesquisa em colunas específicas e reduzir o tempo necessário para executar consultas complexas.

Ambos os tipos de bancos de dados podem se beneficiar da indexação, e a escolha de uma abordagem específica dependerá das necessidades específicas de cada projeto. No geral, a indexação em bancos de dados

relacionais é mais flexível e pode ser aplicada a diferentes tipos de dados, enquanto a indexação em bancos de dados textuais é mais focada em palavras-chave específicas.

- **Pesquisa sobre os dados de forma completa - Query**

O conceito de query pode ser definido como uma consulta a um banco de dados. Dito isso, os bancos de dados textuais geralmente são projetados para permitir a pesquisa rápida de texto, em vez de recuperar dados estruturados como em bancos de dados relacionais. Ou seja, os usuários geralmente pesquisam por palavras-chave ou frases específicas (WITTEN; MOFFAT; BELL, 1999, p. 153), em vez de usar consultas complexas com várias condições de pesquisa. Além disso, os bancos de dados textuais podem usar técnicas como a indexação de texto para acelerar a pesquisa.

Já em bancos de dados relacionais, uma query envolve o uso da linguagem SQL (Structured Query Language) para recuperar dados de uma ou mais tabelas. O SQL permite que os usuários especifiquem as condições de pesquisa que desejam aplicar aos dados, como as colunas que desejam recuperar, as tabelas que desejam pesquisar e as condições de filtro para restringir os resultados. Os bancos de dados relacionais também permitem a realização de operações matemáticas e de agregação em dados, como soma, média e contagem.

E) Desvantagens da Tecnologia Pesquisada

- **Consumo de espaço de armazenamento**

O consumo de armazenamento é uma das principais desvantagens dos bancos de dados textuais. Isso ocorre porque esses bancos de dados são projetados para armazenar grandes quantidades de texto, e o texto pode ocupar muito espaço em disco mesmo utilizando as técnicas de compressão. Ao contrário dos bancos de dados estruturados, que armazenam dados em tabelas com colunas e linhas, os bancos de dados textuais armazenam informações em blocos de texto que são indexados e pesquisados por meio de palavras-chave. Isso pode levar a um aumento significativo no tamanho do

arquivo de banco de dados, especialmente se o texto contido no banco de dados for muito longo.

O armazenamento de dados em bancos de dados textuais pode ser ainda mais desafiador se o banco de dados contiver vários idiomas, pois diferentes idiomas podem exigir diferentes tipos de codificação de caracteres, o que pode afetar o tamanho do arquivo de banco de dados.

- **Complexidade na modelagem de dados**

A modelagem de dados em bancos de dados textuais pode ser mais complexa, pois as informações podem ser apresentadas em diferentes formatos, como documentos, e-mails, mensagens de texto e posts de mídia social. É necessário um esforço adicional para garantir que esses diferentes tipos de dados possam ser integrados e relacionados uns aos outros.

A modelagem de dados em um banco de dados relacional é relativamente simples, pois é baseada em tabelas com colunas e restrições de integridade de dados. Os dados são normalizados para evitar redundância e manter a consistência dos dados. Isso torna a modelagem de dados em bancos de dados relacionais mais fácil de entender e gerenciar.

- **Dificuldade na detecção de erros**

Como os bancos de dados textuais são altamente dependentes de técnicas de processamento de linguagem natural, eles estão sujeitos a erros, como a classificação incorreta de palavras-chave, reconhecimento de caracteres errados e segmentação de palavras erradas. A detecção desses erros pode ser difícil e requer treinamento especializado.

No caso dos bancos de dados relacionais, como as informações são armazenadas em tabelas com colunas definidas e tipos de dados específicos, é possível aplicar restrições de integridade de dados e validação de entrada para detectar erros em tempo real. Além disso, os erros geralmente ocorrem em campos específicos da tabela, tornando mais fácil identificá-los e corrigi-los.

- **Falta de padronização**

A falta de padronização em bancos de dados textuais pode levar a inconsistências e erros de interpretação, especialmente quando várias pessoas

estão inserindo dados no banco de dados. Isso pode afetar a precisão e a qualidade dos resultados obtidos a partir do banco de dados.

Enquanto que nos bancos de dados relacionais a padronização é uma das principais vantagens. Os bancos de dados relacionais são projetados para armazenar dados estruturados em tabelas, onde as colunas têm um tipo de dados específico e são nomeadas de forma consistente. Isso permite que as informações sejam facilmente comparadas e relacionadas, garantindo a integridade dos dados. Além disso, os bancos de dados relacionais seguem um padrão de linguagem de consulta chamado SQL (Structured Query Language), o que torna possível escrever consultas que podem ser executadas em qualquer banco de dados relacional que suporte SQL. Isso proporciona uma grande flexibilidade na escolha do software de gerenciamento de banco de dados e ajuda a garantir a portabilidade dos dados.

F) Exemplo de uso

Existem diversas opções de bancos de dados textuais disponíveis no mercado, mas dois exemplos populares são o Elasticsearch e o Apache Solr. Ambos são de código aberto e podem ser usados gratuitamente, sem custos de licença. O Elasticsearch é construído sobre o Apache Lucene e oferece recursos avançados de pesquisa em texto, incluindo pesquisa de texto completo, sugestões de texto e análise de texto. Já o Apache Solr é uma plataforma de busca em texto completa que permite indexação e busca rápida de grandes volumes de dados não estruturados. Ambas as ferramentas são amplamente utilizadas em ambientes corporativos e em projetos de código aberto para implementação de soluções de busca e análise de texto.

Devido às vantagens que os bancos de dados textuais oferecem, cada vez mais empresas estão adotando essa tecnologia para aprimorar suas operações e melhorar a experiência do usuário. Com bancos de dados textuais, é possível indexar e pesquisar informações de maneira mais rápida e eficiente, permitindo que as empresas processem grandes quantidades de dados e ofereçam respostas mais precisas e personalizadas para seus usuários. Além disso, esses bancos de dados também permitem que as

empresas analisem dados não estruturados, como texto e mídia social, o que pode fornecer informações valiosas sobre sentimentos, tendências e opiniões dos clientes. Dessa forma, as empresas podem melhorar a tomada de decisões, otimizar suas estratégias de marketing e melhorar a qualidade do serviço que oferecem. Abaixo algumas empresas que utilizam essa tecnologia:

- **Airbnb**: utiliza o Elasticsearch para indexar e pesquisar milhões de listagens em tempo real, facilitando aos usuários a busca por acomodações específicas em todo o mundo.
- **The New York Times**: utiliza o Elasticsearch para indexar e pesquisar todos os seus artigos e notícias, permitindo que os leitores pesquisem e encontrem facilmente conteúdo específico em seu site e aplicativo móvel. Além disso, eles também usam o Solr para indexar e pesquisar informações do arquivo histórico de mais de 150 anos de jornalismo.
- **Netflix**: utiliza bancos de dados textuais para melhorar sua recomendação de conteúdo aos usuários. Eles coletam dados sobre as escolhas de visualização e as classificações dos usuários, bem como informações sobre o conteúdo em si. Eles usam técnicas de aprendizado de máquina para analisar esses dados e fazer recomendações de conteúdo personalizado para cada usuário, o que tem sido um grande fator de sucesso para a empresa.
- **Google:** usam tecnologia de processamento de linguagem natural para entender o significado do texto em pesquisas e fornecer resultados relevantes para os usuários. Além disso, eles usam bancos de dados textuais para melhorar o desempenho do Google Translate, que traduz automaticamente textos em diferentes idiomas.

G) Referências Bibliográficas

NAVARRO, G. (2005). Text Databases. In L. Rivero, J. Doorn, & V. Ferraggine (Eds.), Encyclopedia of Database Technologies and Applications (pp. 688-694). Idea Group Inc.Pennsylvania, USA.

ZOBEL, Justin; MOFFAT, Alistair; SACKS-DAVIS, Ron. **An efficient indexing technique for full-text database systems**. In: Proceedings of the International Conference on Very Large Data Bases. p. 352, 1992.

WITTEN, Ian H.; MOFFAT, Alistair; BELL, Timothy C. **Managing Gigabytes: Compressing and Indexing Documents and Images**, Second Edition. San Francisco: Morgan Kaufmann Publishers, 1999.

ROSA, Janaina. **Um Estudo de Compactação de Dados para Biossequências**. 2006. Tese (Programa de Pós-Graduação em Informática) - PUC-RJ, [*S. l.*], 2006. p. 45. Disponível em: https://www.maxwell.vrac.puc-rio.br/9762/9762_5.PDF. Acesso em: 15 abr. 2023.

GOOGLE. **Guia detalhado sobre como a Pesquisa Google funciona**. Disponível em:https://developers.google.com/search/docs/fundamentals/how-search-works?hl=pt-br. Acesso em 16 abr. 2023.

HANNA, Katie. **What is unstructured data?**. Disponível em: https://www.techtarget.com/searchbusinessanalytics/definition/unstructured-data#:~:text=Retailers%2C%20manufacturers%20and%20other%20companies,customer%20service%20and%20corporate%20brands. Acesso em: 16 abr. 2023.

HUTTER, Alex; JHAVERI, Falguni; SAYEEBABA, Senthil. **How Netflix Content Engineering makes a federated graph searchable**. Disponível em: https://netflixtechblog.com/how-netflix-content-engineering-makes-a-federated-graph-searchable-5c0c1c7d7eaf. Acesso em: