UNIVERSIDADE DE BRASÍLIA

Faculdade do Gama

Sistemas de Banco de Dados 2

Tecnologias de Banco de Dados (TI-BD)

Banco de Dados Textuais

Nome: Felipe Borges de Souza Chaves

Matrícula: 16/0049733

Brasília, DF

2019

1. Introdução

Com o crescimento de informação não estruturada na web e com um movimento de informatização massivo feito por instituições públicas como cartórios, bibliotecas, tribunais, etc; as tecnologias de de rankeamento e busca textual ficaram bem famosas. Diferentemente de uma base de dados normais o foco dessas tecnologias e buscar informação em textos e recuperar conteúdo relevante para seu usuário. Em função dessa característica majoritária em busca textual databases de textos também são referenciadas como sistema de recuperação de informação ou sistema de texto completo.

2. Definição da Tecnologia Pesquisada

É importante notar que a informação textual pode ser tratada de maneira não estruturada, ou seja, bancos de dados especializados em textos podem ser relacionais ou não. Além disso, esse tipo de tecnologia geralmente tem suporte para vários tipos de documentos (.doc, .docx. .pdf, .xml, etc) para que se possa recuperar informações facilmente de documentos institucionais ou de conjunto de dados históricos do usuário.

Para realizar a otimização na busca os bancos de dados textuais utilizam várias tecnologias de databases não estruturadas, alguns algoritmos de persistências alternativos a B^+ tree e outros algoritmos comuns a databases convencionais e principalmente técnicas de indexação próprias para solução do problema em busca de coleções de documentos.

Por fim, a recuperação de informação em texto é uma área tão importante que existe uma conferência apenas para o assunto: **Text REtrival Conference (TREC)**.

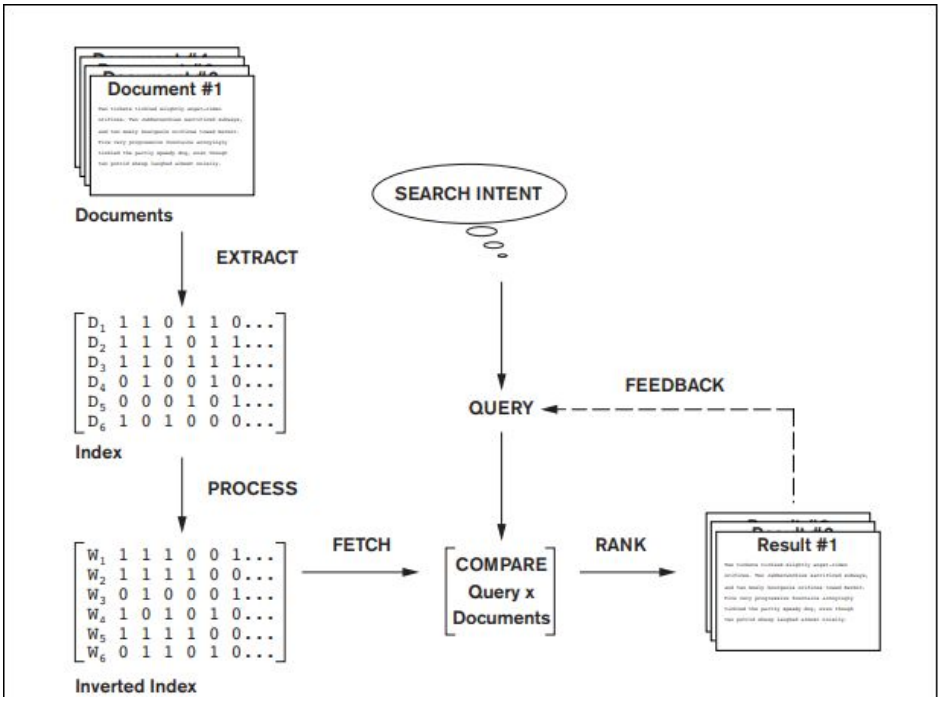

Imagem 1: Ilustração do processo de busca em database textuais. Retirado de **Fundamental of database systems[5].**

3. Objetivo principais da tecnologia pesquisada

- Busca em coleções de documentos e textos;
- Rankeamento de conteúdo relacionado com a pesquisa;
- Otimização de buscas em coleções de dados(textos) não estruturados;

4. Comparação entre tecnologias

É Importante ressaltar que databases de textos tem um propósito específico como ressaltado na introdução da tecnologia, portanto, nessa comparação estaremos apenas analisando as vantagens e desvantagens no quesito de textos. Além disso, comparar as tecnologias de recuperação de textos com outras no mercado é uma tarefa difícil pois as tecnologias especializadas em textos lançam mão de outras tecnologias difundidas no mercado como o NoSQL.

5. Vantagens da tecnologia

- Extremamente eficiente na busca de coleções de textos não estruturados, sendo uma tecnologia importante para buscar informações em textos não estruturados.
- Uma query nesta tecnologia não necessariamente retorna o conteúdo e sim apenas referência para eles, dessa forma otimizando a busca em coleções gigantes de dados.
- Além da busca recuperar resultados que se encaixam na pesquisa os bancos de textos também conseguem trazer dados que são próximos ao conteúdo buscado, usando técnicas de rankeamento para mapear a relevância de cada documento para aquela busca.

6. Desvantagens da tecnologia

- Não existe um padrão de linguagens de pesquisa, como o SQL. Cada mecanismo de recuperação de informação implementa um ou vários métodos como: modelo booleano, vetores no espaço ou técnicas probabilísticas.
- O armazenamento pode ser ligeiramente mais dispendioso do que um mecanismo de banco de dados normais, visto que a indexação inversa é feita a partir de coleções de palavras e não chaves.
- Para documentos com formatos padrões essas técnicas podem não ser a melhor alternativa, sistemas de textos completos partem do suposição que os documentos que estão sendo salvos não tem características relacionais, dessa forma os cálculos e formas de indexação disponibilizadas pelos mesmo podem não ser a melhor alternativa em casos onde o formato do arquivo e a estrutura não varie tanto.

7. Modelos de busca em bancos textuais

A necessidade da busca em coleções de dados textuais faz com que as tecnologias de sistemas de bancos usem várias algoritmos alternativos à lógica relacional em bancos relacionais comuns, como exposto na secção de desvantagens estas tecnologias não tem um padrão de sistema de busca. Nesta secção vamos conversar um pouco sobre três métodos para recuperação de dados e um método muito utilizado para indexação do mesmo.

7.1. Busca booleana

A busca booleana consiste usar a lógica dos conjuntos (a mesma da lógica relacional) para elaborar um conjunto resultado da pesquisa. Neste tipo de busca os operadores AND, OR, NOR são utilizados para relacionar palavras da pesquisa e fazer com o conjunto de dados fique mais específico.

Além de palavras as buscas booleanas podem ser utilizados em cima dos metadados do arquivo, como por exemplo: autor, data de publicação, instituições dentre outros. A principal dificuldade existente nesse mecanismo é que o conjunto resposta é exatamente o que se adere a busca. Sendo assim, não existem mecanismos de ranqueamento ou busca por similaridade.

7.2. Modelo de vetores no espaço

Em linhas bem gerais essa técnica propõe colocar cada documento como um vetor em um espaço n-dimensional retirando algumas características do texto. A busca também é considerada um vetor n-dimensional que se encontra no mesmo espaço. A finalidade dessa abordagem é encontrar uma grandeza de distância entre estes vetores, assim essas buscas são capazes de trazer não só o conteúdo que combina com a busca mas também o conteúdo que está relacionado com ela.

A principal vantagem dessa técnica e ranquear os documentos mais próximos à busca usando algoritmos de remoção de características do texto. A grande desvantagem é o processamento, para realizar o ranqueamento cada documento deve passar por um pré-processamento o que pode sobrecarregar o computador onde a base está alocada.

7.3. Modelos Probabilísticos

Este modelo também busca ranquear os documentos (ou parte deles) a uma determinada busca, porém a abordagem aqui utiliza de métodos estatísticos para realizar esse relacionamento. Essa classe de algoritmos de busca é muito utilizado em engine de buscas por poder relacionar não somente a busca do usuário atual mas também relacionar a busca de vários usuários a um único tópico.

A ideia por trás desses algoritmos é a categorização da busca em um conjunto de textos relevantes ou não relevantes. Desta forma conseguimos movimentar documentos de um conjunto para outro de forma a diminuir o erro inerente a busca no sistema.

7.4. Indexação invertida

Para que todos os algoritmos anteriores tenham performance é necessário ter uma mecanismos eficiente de indexação dos documentos. Dessa forma surgiu o conceito de indexação invertida. Mais objetivamente segundo [5]:

"Um índice invertido de uma coleção de documentos e uma estrutura de dados que relaciona termos distintos com uma lista de todos os documentos que contém o mesmo" - Fudamental of database systems.

Como já deve ter ficado claro nas outras secções um sistema de banco textuais busca performance em buscas, dessa forma é razoável pensar que a indexação dos arquivos seria realizado pelas informações presente no conteúdo do texto. Algumas características não tão triviais dessa indexação é que o banco não realiza produtos cartesianos entre tabelas. A indexação invertida tem a sua própria maneira de realizar relações entre seus índices.

8. Aplicações de bancos textuais

Como já foi exposto os bancos textuais são importantes na área de recuperação de informação. Algumas instituições como bibliotecas, tribunais e governos utilizam essa tecnologia para extrair informação de dados históricos. Um advogado por exemplo pode utilizar dessa tecnologia para encontrar todos os processos relacionados com uma busca e assim se preparar melhor um caso.

Outra grande aplicação de banco de dados textuais (diria que até massiva) é na área da genética. Estes pesquisadores estão constantemente tentando relacionar partes do nosso código genético com código de outras espécies para realizar o mapeamento evolutivo.

Para os profissionais de tecnologia estamos o tempo todo utilizando bancos textuais. Não somente para busca mas também para identificar rapidamente qual foi o problema que um sistema encontrou por meio dos arquivos de logs ou até mesmo buscando uma parcela de código no meio de milhares de linhas.

9. Principais tecnologias

Atualmente as melhores tecnologias de bancos textuais são o Elastic Search, que tem um secção do software open source e oferecem cursos para a plataforma Elastic Search, em

contrapartida, existem soluções como o Loki da organização Grafana que é puramente open source e mais leve para computadores mais fracos.

Empresas como Netflix e Google estão utilizam essa tecnologia para conseguir encontrar informações em vários computadore espalhados pelo mundo. Além disso instituições como STF e TCU também começaram a utilizar a OSS dessas empresas e portanto utilizam dessas tecnologias para acompanhamento de aplicações em tempo real em clusters kubernetes.

10. Conclusão

Base de dados especializadas em textos são extremamentes relevantes para a humanidade. Ela está presente em nossas pesquisas do dia-a-dia e até mesmo em nossas pesquisas acadêmicas. Nesse quadro cabe ao profissional de engenharia ter conhecimento e técnica para elaborar novas tecnologias e soluções que atendam a necessidade desses variados tipos de usuários.

Referências

1. Kehoe, Miles. Contrasting Relational Databases and Full-Text Search Engines, 2009. Disponível em: http://www.ideaeng.com/database-full-text-search-0201. Acesso 8 Set. 2019.
2. Tweedie, Mitchell. Types of databases and DMS (with examples). Disponível em: https://codebots.com/data-management/types-of-databases-and-dbms-with-examples. Acesso 8 Set. 2019.
3. ELMASRI, Ramez. **Fundamentals of database systems**. 2017.
4. LESTER, Nicholas; MOFFAT, Alistair; ZOBEL, Justin. Efficient online index construction for text databases. **ACM Transactions on Database Systems (TODS)**, v. 33, n. 3, p. 19, 2008.
5. LOEFFEN, Arjan. Text databases: A survey of text models and systems. **Sigmod record**, v. 23, n. 1, p. 97-106, 1994.