UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CAMPUS DOIS VIZINHOS
CURSO DE ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS

PEDRO HENRIQUE BERGAMO BERTOLLI

CONSTRUÇÃO DE UM BANCO DE DADOS DE EXPERIMENTOS DE ÍNDICES ESPACIAIS PARA TREINAMENTO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA

TRABALHO DE CONCLUSÃO DE CURSO DE ESPECIALIZAÇÃO

DOIS VIZINHOS
2022

PEDRO HENRIQUE BERGAMO BERTOLLI

CONSTRUÇÃO DE UM BANCO DE DADOS DE EXPERIMENTOS DE ÍNDICES ESPACIAIS PARA TREINAMENTO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA

Trabalho de Conclusão de Curso de Especialização apresentado ao Curso de Especialização em Ciência de Dados da Universidade Tecnológica Federal do Paraná, como requisito para a obtenção do título de Especialista em Ciência de Dados.

Orientador: Prof. Dr. Yuri Kaszubowski Lopes
Coorientador: Prof. Dr. Anderson Carniel Chaves

DOIS VIZINHOS
2022

PEDRO HENRIQUE BERGAMO BERTOLLI

CONSTRUÇÃO DE UM BANCO DE DADOS DE EXPERIMENTOS DE ÍNDICES ESPACIAIS PARA TREINAMENTO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA

Trabalho de Conclusão de Curso de Especialização apresentado ao Curso de Especialização em Ciência de Dados da Universidade Tecnológica Federal do Paraná, como requisito para a obtenção do título de Especialista em Ciência de Dados.

Data de aprovação: 11/novembro/2022

Yuri Kaszubowski Lopes
Doutorado
Universidade do Estado de Santa Catarina

Rafael Alves Paes de Oliveira
Doutorado
Universidade Tecnológica Federal do Paraná - Câmpus Dois Vizinhos

Francisco Carlos Monteiro Souza
Doutorado
Universidade Tecnológica Federal do Paraná - Câmpus Dois Vizinhos

DOIS VIZINHOS
2022

AGRADECIMENTOS

Primeramente agradeço a minha esposa Francine e minha filha Clara por sempre estrarem ao meu lado ao longo desta jornada.

Agradeço a UTFPR-DV por disponibilizar esse curso de especialização inteiramente à distância, o que me permitiu cursá-lo assim como quase todos colegas de classe aos quais tive o prazer de conhecer durante o curso.

Agradeço ao Anderson Chaves Carniel pelo período em que foi meu orientador, pela paciência e dedicação em minha orientação, projetos e momentos compartilhados.

Agradeço também ao meu orientador Yuri Kaszubowski Lopes pelo apoio e orientação deste trabalho.

RESUMO

Bancos de dados espaciais vem sendo amplamente utilizados em diversas aplicações de nosso cotidiano, desde serviços baseados em localização à sistemas de agricultura de precisão. O crescente volume de dados gerados e consumidos por essas aplicações e aliados à complexidade existente em manipular esse tipo de informação torna necessário o estudo de técnicas de otimização para que a recuperação desses dados seja feita da maneira mais eficiente possível. Uma das técnicas mais utilizadas é a indexação espacial que tem como principal objetivo reduzir o espaço de busca e consequentemente o processamento computacional. Porém existem diversos índices na literatura e diferentes tipos de configurações à eles associadas, fazendo com que um determinado índice combinado a uma determinada parametrização obtenha melhor desempenho sobre uma base de dados. Nesse sentido, o objetivo desse trabalho é realizar testes experimentais para coletar informações estatíticas de bancos de dados indexados e construir um conjunto de dados com características necessárias para que um algoritmo de aprendizado de máquina seja capaz de ser treinado e prever o melhor índice e suas configurações para uma determinada base de dados.

Palavras-chave: bancos de dados espaciais; indexação; testes de desempenho; coleta de informações.

ABSTRACT

Spatial databases have been widely used in myriad applications of our daily lives, from location-based services to precision agriculture systems. The growing volume of data generated and consumed by these applications and combined to the existing complexity in handling this type of information makes it necessary to study optimization techniques to query this data in the most efficient way possible. One of the most used techniques is spatial indexing, whose main objective is to reduce the search space and consequently the computational processing. However, there are several indices in the literature and different types of configurations associated with them, making a certain index combined with a certain parametrization obtain better performance than others on a database. In this sense, the objective of this work is to execute experimental tests to collect statistical information from indexed databases and build a dataset with characteristics necessary for a machine learning algorithm that will be able to be trained and predict the best index and its settings for a given spatial database.

Keywords: spatial databases; indexing; performance tests; information extraction.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A ascensão e popularização da Internet das Coisas (IoT), consequência direta da Indústria 4.0, conferiu à localização um papel bastante importante no cotidiano das pessoas. A localização exata de um objeto ou indivíduo permite identificar distâncias e relações espaciais, o que pode ser altamente relevante no planejamento de rotas, monitoramento de desastres naturais e até na simulação e mapeamento de ambientes.

A ciência de dados espaciais trata justamente a localização, a distância e a interação espacial como aspectos centrais dos dados e emprega métodos para armazenar, visualizar, interpretar e consultar esses dados (CARNIEL; SCHNEIDER, 2021). No que se refere ao processamento de consultas espaciais, sistemas de informação geográfica (GIS) e sistemas de banco de dados espaciais (SDBS) são comumente utilizados. Tais sistemas são responsáveis por gerenciar objetos de tipos espaciais, sendo esses representados por atributos geométricos, como pontos, linhas e regiões (polígonos).

A recuperação de objetos espaciais vem sendo amplamente abordada na literatura nos últimos anos, envolvendo estudos em diversas frentes (CARNIEL, 2020). Dentre as mais desafiadoras, tem-se o desenvolvimento de técnicas de indexação eficiente.

Diversos índices (GUTTMAN, 1984; BECKMANN et al., 1990; KAMEL; FALOUTSOS, 1994; CARNIEL; CIFERRI; CIFERRI, 2016; CARNIEL; CIFERRI; CIFERRI, 2017) foram propostos na literatura pela necessidade de recuperar objetos espaciais em curto espaço de tempo. No entanto, cada índice espacial contém seus próprios conjuntos de parâmetros que afetam diretamente o seu desempenho no conjunto de dados que ele indexa. Desse modo, um índice espacial pode assumir diferentes configurações. Esse número de configurações pode ser bem alto, dificultando a escolha do melhor índice espacial e seus parâmetros para responder uma determinada consulta espacial.

Nesse sentido, o desenvolvimento de um método de aprendizado de máquina capaz de determinar o melhor índice e os melhores parâmetros a ele associados para processar uma determinada consulta seria uma das opções para resolver esse problema. Porém, a etapa que antecede a concepção desse método é algo de extrema importância para que ele seja realizado: a condução de testes de desempenho para a extração de um conjunto de características (*features*) que descreva as características de execuções de diferentes tipos de índices e consultas.

Desse modo, o objetivo principal desse trabalho é a construção de um conjunto de features geradas a partir de exaustivas cargas de trabalho, com o auxilio do framework FESTIval (CARNIEL; CIFERRI; CIFERRI, 2020), realizadas em bases de dados espaciais (reais e sintéticas).

1.1 Problema de Pesquisa

Quando o assunto aprendizado de máquina é colocado em discussão, logo são colocados em jogo os algoritmos e métodos que serão utilizados para satisfazer determinado problema. Porém, deixamos de lado uma etapa importante de todo esse processo que é a coleta e construção da base de informação que será utilizada pelo algoritmo em seu treinamento.

Particularmente no caso deste trabalho, serão executadas três etapas:

1. Extração de dados espaciais reais e sintéticos;
2. Realização de experimentos baseados na execução de operações utilizando índices espaciais;
3. Coleta de informações estatísticas referentes às operações executadas e as características específicas de cada conjunto de dados indexado.

1.2 Objetivos

Considerando a importância de escolher o melhor tipo de configuração de um índice espacial e a dificuldade em realizar experimentos para a coleta desse tipo de informação, os principais objetivos do trabalho são apresentados a seguir.

1.2.1 Objetivo Geral

Considerando o contexto e problema que o envolve, o objetivo geral deste trabalho é a construção de um conjunto de features baseadas em operações de bancos de dados espaciais indexados. Neste trabalho, o conjunto de features se refere as características descritivas das bases de dados com as características de seus índices e estatísticas de execuções.

1.2.2 Objetivos Específicos

Para que o objetivo principal seja atendido será necessário conduzir testes experimentais para coletar informações de desempenho dos índices. Estes teste serão realizados em conjuntos de dados reais e sintéticos para que a variação dos objetos espaciais seja atendido. A avaliação de desempenho será realizada através do framework FESTIval (CARNIEL; CIFERRI; CIFERRI, 2020).

1.3 Justificativa

Muitos trabalhos que fazem o uso de inteligência artificial ou análise exploratória de dados raramente disponibilizam a maneira em que o conjunto de dados de treinamento foi construído, pois colocam o foco nos métodos e algoritmos utilizados.

Um dos trabalhos relacionados (MARCUS et al., 2019) descreve como importantes características gerais referentes ao conjunto de dados espaciais foram extraídas, como, por exemplo, a área média das aproximações geométricas. Porém, não são relatados detalhes de

como a coleta de informações relacionadas aos índices foi realizada, tais como características de configuração e execução de operações.

Desse modo, este trabalho visa contribuir com lacunas encontradas na literatura no que se refere aos experimentos necessários para a construção de um conjunto de features visando disponibilizá-lo para posterior implementação de métodos preditivos ou análise exploratória.

1.4 Materiais e Métodos

Os dados espacias reais serão extraídos via OpenStreetMap (OpenStreetMap contributors, 2017) e carregados diretamente em um banco de dados postgreSQL enquanto dados sintéticos serão gerados a partir do framework Spider (KATIYAR et al., 2020) em formato csv.

O framework FESTIval (CARNIEL; CIFERRI; CIFERRI, 2020) será utilizado na avaliação de desempenho dos índices espaciais para coletar informações estatísticas das cargas de trabalho executadas.

Por fim, todas as características extraídas ao longo desse processo serão unificadas em um apenas um conjunto de dados que irá conter as seguintes informações: características do conjunto de dados, informação sobre a estrutura do índice, configuração única de um índice e informações estatísticas sobre a execução de operações usando o índice.

1.5 Organização do Trabalho

Como base para o entendimento do trabalho apresentado, o Capítulo 2 apresenta uma breve introdução sobre bancos de dados espaciais e seus tipos de dados. O Capítulo 3 apresenta como as manipulações de dados espaciais são realizadas através de relacionamentos topológicos. O processamento de consultas com a utilização de aproximações geométricas é abordado no Capítulo 4 e a revisão da literatura se encerra com a descrição e uso de índices espaciais no Capítulo 5.

O Capítulo 6 apresenta a forma de como os materiais e métodos foram utilizados no desenvolvimento desse trabalho assim como os resultados obtidos. Por fim, as considerações finais estão presentes no Capítulo 7.

2 BANCOS DE DADOS ESPACIAIS

Bancos de dados espaciais, também conhecidos como banco de dados geográficos, são bancos de dados capazes de suportar o armazenamento e manipulação de objetos espaciais. Os objetos espaciais utilizam um tipo de dado geométrico representado em um espaço Euclidiano, podendo possuir 0,1 e 2 dimensões como por exemplo pontos, linhas e polígonos respectivamente.

Esses objetos ainda podem ser caracterizados em objetos simples ou compostos. O que os diferencia é a quantidade de componentes que cada um pode conter, sendo que objetos simples apresentam apenas um componente enquanto objetos compostos apresentam um número finito deles. Por exemplo, podemos considerar a região da Oceania como um objeto complexo onde cada ilha que compõe o continente é considerada um componente e o conjunto desses componentes é responsável por definir todo território. A Figura 1 a seguir ilustra esses tipos de dados.

Figura 1 – Exemplo de tipos de objetos espaciais: (a) ponto simples; (b) ponto complexo;(c) linha simples; (d) linha complexa; (e) região simples e (f) região complexa.

Fonte: Autoria Própria

3 RELACIONAMENTOS TOPOLÓGICOS

A manipulação de objetos espaciais é realizada através de relacionamentos, sendo possível defini-los em relacionamentos espaciais dos tipos: métricos, direcionais e topológicos. Os relacionamentos métricos referem-se à distância entre objetos enquanto os direcionais consideram a direção para satisfazer uma determinada condição. Por exemplo, é possível responder perguntas similares a: "Quais são os países que se encontram ao sul da linha do Equador ?". Os relacionamentos topológicos são definidos pela combinação de interseções entre objetos levando em conta seu interior, borda e exterior. Esse conjunto de combinações forma uma matriz chamada de Modelo de 9-Interseções (*9-IM)* (EGENHOFER; HERRING, 1990), em que cada célula contém um valor booleano indicando se aquela interseção existe (*True*) ou não (*False*).

Figura 2 – Exemplo do modelo de configuração de uma matriz 9-IM

$$\mathrm{9IM}(a,b) = \begin{bmatrix} a^o \cap b^o \neq \emptyset & a^o \cap \partial b \neq \emptyset & a^o \cap b^e \neq \emptyset \\ \partial a \cap b^o \neq \emptyset & \partial a \cap \partial b \neq \emptyset & \partial a \cap b^e \neq \emptyset \\ a^e \cap b^o \neq \emptyset & a^e \cap \partial b \neq \emptyset & a^e \cap b^e \neq \emptyset \end{bmatrix}$$

Fonte: Autor Desconhecido https://en.wikipedia.org/wiki/DE-9IM

Destacam-se alguns tipos de consultas espaciais (CARNIEL, 2020) que satisfazem um determinado relacionamento topológico, dentre elas estão a *point query* (PQ), *containment range query* (CRQ) e *range query* (RQ). As consultas do tipo PQ selecionam todos os objetos que contém dentro de sua geometria um determinado ponto passado como parâmetro. As consultas RQ selecionam todos os objetos que obedecem um tipo de relacionamento topológico em relação a uma determinada região chamada de janela de consulta (*window query* - WQ). Uma CRQ é uma RQ que aplica o relacionamento topológico contém (*contains*), ou seja, seleciona todos os objetos espaciais que estão contidos em uma janela de consulta.

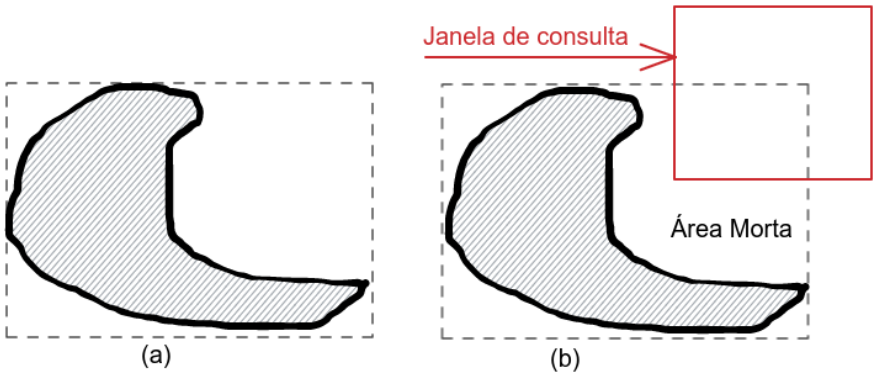
4 PROCESSAMENTO DE CONSULTAS

Visto que cada configuração da matriz 9-IM corresponde a um tipo de relacionamento, o processamento dessa matriz torna-se muito custoso, pois envolve complexos algoritmos de geometria computacional para realizar o cálculo das interseções entre objetos.

Com o objetivo de reduzir a complexidade dessas operações, aproximações geométricas são aplicadas aos objetos espaciais, tornando possível representá-los com apenas dois pares de coordenadas. Uma representação comumente utilizada é o MBR (*Minimum Bounding Rectangle)* (BERTELLA et al., 2021), que representa o menor retângulo possível que contém todos os pontos de um objeto espacial.

Porém, essas aproximações podem conter imprecisões em relação a verdadeira geométrica do objeto e retornar respostas falso positivas. A figura Figura 4 ilustra esse caso específico, onde uma janela de consulta intersecta uma área vazia desse MBR, essa tal área recebe o nome de área morta(*dead space*).

Figura 3 – a) MBR de um objeto espacial e sua área morta (b) retornando um resultado falso positivo



Fonte: Autoria Própria

Para solucionar esse problema, duas etapas são realizadas no processamento de consultas espaciais:

- Etapa de Filtragem: Candidatos a resposta são selecionados de acordo com sua aproximação geométrica.
- Etapa de Refinamento: Para cada candidatos da etapa anterior a geometria exata é acessada e verificado se satisfaz o predicado topológico da consulta.

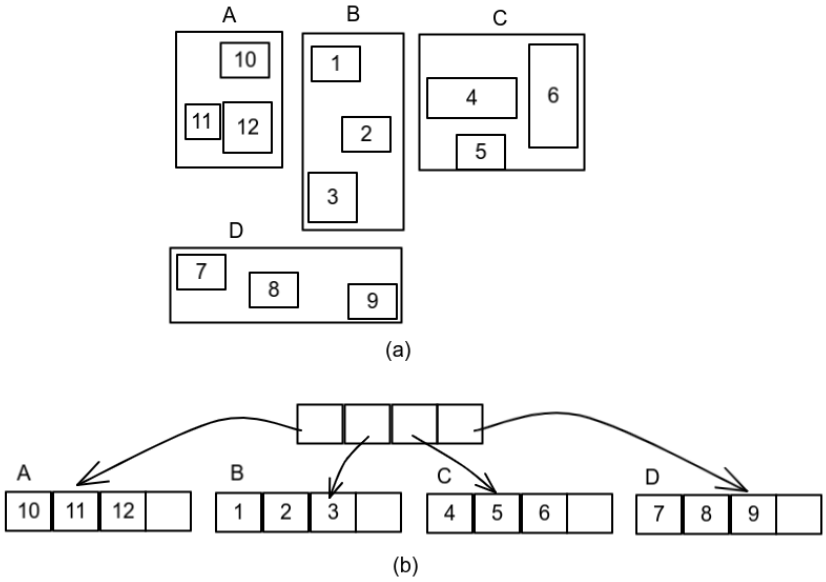
5 ÍNDICES ESPACIAIS

Os índices espaciais são estruturas de dados hierárquicas que armazenam aproximações geométricas, e portanto, permitem o acesso eficiente a objetos espaciais, reduzindo o espaço de busca e, consequentemente, o custo de processamento computacional.

Um dos índices espaciais mais conhecidos na literatura é a R-*Tree* (GUTTMAN, 1984), essa estrutura de dados armazena as aproximações geométricas dos objetos espaciais em uma árvore balanceada. Esse mesmo conceito também é utilizado por suas implementações variantes, tais como a R*-*Tree* (BECKMANN et al., 1990), R^+ (SELLIS; ROUSSOPOULOS; FALOUTSOS, 1987) e Hilbert R-*Tree* (KAMEL; FALOUTSOS, 1994).

Estruturalmente, cada nó da árvore corresponde a uma página de disco composto por no máximo M entradas e no mínimo m, onde m $<=$ M/2. Um nó de uma R-tree pode possuir dois tipos, nós internos e nós folhas onde ambos podem ser representados pela forma (mbr, ptr). Para os nós internos o mbr corresponde ao MBR que contém todos MBRs armazenados nas entradas de seu nó filho e ptr é o ponteiro para este nó, diferentemente do nó folha onde através do ponteiro é possível acessar diretamente a geometria exata de um objeto espacial.

Figura 4 – Representação gráfica (a) e hierárquica (b) de uma R-tree.



Fonte: Autoria Própria

6 MATERIAIS E MÉTODOS

Este capítulo está organizado da seguinte forma: Na Seção 6.1 é descrito como os dados foram extraídos para dar início aos trabalhos. Na Seção 6.2 são apresentados os experimentos realizados no framework FESTIval e o produto final obtido em Seção 6.3.

6.1 Extração de Dados Reais e Sintéticos

A construção do dataset utilizado no treinamento do modelo de machine learning é composta por dados sintéticos e reais com o objetivo de variar a distribuição e complexidade dos objetos espaciais.

Os dados sintéticos foram preparados através do gerador Spider (KATIYAR et al., 2020), onde 22 conjuntos de dados foram gerados com diferentes características, tais como o tipo de distribuição dos objetos espaciais (uniform, diagonal, gaussian, sierpinski, bit e parcel) e o tipo de geometria (ponto ou retângulo). Os dados reais foram extraídos a partir do OpenStreetMap contendo 6 datasets divididos em três conjuntos correspondentes a cada tipo de dado espacial utilizado (ponto, linha ou polígono). Os detalhes destes conjuntos de dados são fornecidos na Tabela 1:

Tabela 1 – Informações dosbre os dados extraídos

Tabela	Tipo de Dado	Volumetria	Fonte	Descrição
brazil_lakes2022	POLYGON	17998	Geofabrik Download Server	Contém todos os lados do Brasil
brazil_tourism2022	POINT	20354	Geofabrik Download Server	Contém todos os pontos turisticos do Brasil
brazil_railways2022	LINE	24178	Geofabrik Download Server	Contém todos os trilhos de trem do brasil
brazil_natural2022	POINT	308224	Geofabrik Download Server	Contém todas formações naturais do Brasil
brazil_trackroads2022	LINE	276533	Geofabrik Download Server	Contém todas as estradas de terra do Brasil
brazil_leisure2022	POLYGON	223039	Geofabrik Download Server	Contém todas as áreas de laser do Brasil
synthetic_uniform_points_20k	POINT	20000	Spider Data Generator	Contém pontos de acordo com a distribuição uniforme
synthetic_diagonal_points_20k	POINT	20000	Spider Data Generator	Contém pontos de acordo com a distribuição diagonal
synthetic_gaussian_points_20k	POINT	20000	Spider Data Generator	Contém pontos de acordo com a distribuição de Gauss
synthetic_sierpinski_points_20k	POINT	20000	Spider Data Generator	Contém pontos de acordo com a distribuição de Sierpinski
synthetic_bit_points_20k	POINT	20000	Spider Data Generator	Contém pontos de acordo com a distribuição bit
synthetic_parcel_box_20k	POLYGON	20000	Spider Data Generator	Contém retângulos de acordo com a distribuição parcel
synthetic_uniform_box_20k	POLYGON	20000	Spider Data Generator	Contém retângulos de acordo com a distribuição uniforme
synthetic_diagonal_box_20k	POLYGON	20000	Spider Data Generator	Contém retângulos de acordo com a distribuição diagonal
synthetic_gaussian_box_20k	POLYGON	20000	Spider Data Generator	Contém retângulos de acordo com a distribuição de Gauss
synthetic_sierpinski_box_20k	POLYGON	20000	Spider Data Generator	Contém retângulos de acordo com a distribuição de Sierpinski
synthetic_bit_box_20k	POLYGON	20000	Spider Data Generator	Contém retângulos de acordo com a distribuição bit
synthetic_uniform_points_200k	POINT	200000	Spider Data Generator	Contém pontos de acordo com a distribuição uniforme
synthetic_diagonal_points_200k	POINT	200000	Spider Data Generator	Contém pontos de acordo com a distribuição diagonal
synthetic_gaussian_points_200k	POINT	200000	Spider Data Generator	Contém pontos de acordo com a distribuição de Gauss
synthetic_sierpinski_points_200k	POINT	200000	Spider Data Generator	Contém pontos de acordo com a distribuição de Sierpinski
synthetic_bit_points_200k	POINT	200000	Spider Data Generator	Contém pontos de acordo com a distribuição bit
synthetic_parcel_box_200k	POLYGON	200000	Spider Data Generator	Contém retângulos de acordo com a distribuição parcel
synthetic_uniform_box_200k	POLYGON	200000	Spider Data Generator	Contém retângulos de acordo com a distribuição uniforme
synthetic_diagonal_box_200k	POLYGON	200000	Spider Data Generator	Contém retângulos de acordo com a distribuição diagonal
synthetic_gaussian_box_200k	POLYGON	200000	Spider Data Generator	Contém retângulos de acordo com a distribuição de Gauss
synthetic_sierpinski_box_200k	POLYGON	200000	Spider Data Generator	Contém retângulos de acordo com a distribuição de Sierpinski
synthetic_bit_box_200k	POLYGON	200000	Spider Data Generator	Contém retângulos de acordo com a distribuição bit

Fonte: Autoria Própria

6.2 Realização de Experimentos Baseados na Execução de Operações Utilizando Índices Espaciais

Uma das dificuldades em realizar testes de desempenho é o grande esforço que deve ser feito para coletar diferentes tipos de medições realizadas em cada teste. Essa tarefa tem o nível de complexidade ainda mais elevado quando necessitamos realizar experimentos em grandes bases de dados e executar operações inúmeras vezes.

O framework FESTIval (CARNIEL; CIFERRI; CIFERRI, 2020) é uma extensão *open-source* do banco de dados postgreSQL e é capaz de realizar o teste de desempenho em índices espaciais. Isso torna possível a execução de diferentes cargas de trabalho e armazenamento dos resultados estatísticos obtidos em um esquema postgreSQL.

O framework FESTIval foi utilizado neste trabalho de modo a auxiliar nos seguintes tópicos:

- **Configuração do Índice.** Foram criadas diferentes tipos de configuração de indexação utilizando os seguintes tipos de índices espaciais: i) R-Tree, ii) R*-Tree e iii) Hilbert-Tree. A R-Tree usa o algoritmo de split quadrático, para R*-Tree foi empregado o método de reinserção de 30% e para a Hilbert-Tree o método de split *2-to-3*.
  Foi variado o tamanho da página do nós de 1MB a 64MB. Para o buffer, o tamanho utilizado foi de 1 MB com o algortimo de substituição LRU.
- **Cargas de trabalho.** Para cada dataset da Tabela 1 foram executadas as seguintes cargas de trabalho: i) construção do índice; (ii) execuções de 1000 point queries (PQ); iii) execuções de 3000 window queries (WQ); iv) execução de 3000 containment range queries (CRQ).
  Uma PQ retorna os pontos que são iguais a um dado ponto enquanto a WQ retorna todos os objetos que intersectam uma determinada janela de consulta. Por sua vez, uma CRQ retorna os objetos contidos em uma janela de consulta retangular, incluindo suas bordas.
  Três tipos diferentes de tamanhos de janelas foram utilizados contendo 0.001%, 0.01% e 0.1% de tamanho em relação a área total do dataset.

6.3 Extração de Features

Como resultado dos items anteriores Seção 6.1 e Seção 6.2, foi possível construir o conjunto de dados de treino para um modelo de aprendizado de máquina, sendo que cada linha é uma representação da média de execução das operações executadas utilizando uma única configuração de índice.

O conjunto de características desse conjunto de dados de treino é composto por quatro grupos: (i) informações estatísticas sobre o dataset; (ii) informações sobre a estrutura dos índices criados; (iii) parametrização dos índices e (iv) informações estatísticas sobre a execução de operações usando o índice.

Informações estatísticas do dataset. Esse grupo descreve cada conjunto de dados indexado e contém 26 atributos divididos entre informações básicas de identificação do dataset, tais como nome, número de linhas e tipo de geometria, e atributos estatísticos referentes aos objetos espaciais, como, por exemplo, valores mínimos, máximos e médias de propriedades geométricas. A Tabela 2 apresenta os atributos e suas respectivas descrições.

Tabela 2 – Descrição das informações estatísticas do dataset

Grupo	Coluna	Descrição
Informação sobre o dataset	dataset_name	conjunto de dados que foi indexado
	dataset_type	tipo de geometria presente na tabela
	dataset_nrows	total do número de linhas da tabela
	dataset_column_size	tamanho em bytes de toda coluna do tipo geometry da tabela
	dataset_min_area	Menor área de uma geometria encontrada na tabela
	dataset_min_perimeter	Menor perímetro de uma geometria encontrada na tabela
	dataset_min_length	comprimento mínimo de geometria encontrada na tabela
	dataset_min_points	número de pontos da menor geometria da tabela
	dataset_max_area	área máxima de uma geometria encontrada na tabela
	dataset_max_perimeter	perímetro máximo de uma geometria encontrada na tabela
	dataset_max_length	Comprimento máximo de uma geometria na tabela
	dataset_max_points	número de pontos da maior geometria da tabela
	dataset_avg_mbr_area	área média dos mbrs da tabela
	dataset_avg_x_mbr	tamanho médio dos eixos x dos mbrs da tabela
	dataset_avg_y_mbr	tamanho médio dos eixo y dos mbrs da tabela
	dataset_avg_area	área média das geometrias da tabela
	dataset_avg_perimeter	perímetro médio das geometrias da tabela
	dataset_avg_length	comprimento médio das geometrias da tabela
	dataset_avg_points	média do número de pontos das geometrias de uma tabela
	dataset_stddev_area	desvio padrão da área das geometrias de uma tabela
	dataset_stddev_perimeter	desvio padrão do perímetro das geometrias de uma tabela
	dataset_stddev_length	desvio padrão do comprimento das geometrias de uma tabela
	dataset_stddev_points	desvio padrão da quantidade de pontos das geometrias de uma tabela
	dataset_total_area	área total das geomtrias de uma tabela
	dataset_total_perimeter	perímetro total das geometrias de uma tabela
	dataset_total_length	comprimento total das geometrias de uma tabela

Fonte: Autoria Própria

Informações sobre a estrutura dos índices criados. Esse grupo apresenta 9 atributos contendo informações relacionadas à estrutura dos índices utilizados, tais como altura, número de nós e medidas médias encontradas a partir das aproximações geométricas de uma árvore. Esses atributos podem ser conferidos na Tabela 3 abaixo:

Tabela 3 – Descrição das informações sobre a estrutura dos índices criados.

Groupe	Column	Description
Informação sobre a estrutura do índice	height	Altura da árvore
	num_int_nodes	número de nós da árvore
	num_leaf_nodes	número de nós folha da árvore
	num_entries_int_nodes	número de entradas por nó da árvore
	num_entries_leaf_nodes	número de entradas por nó folha da árvore
	avg_num_entries_pnode	média do número de nós da árvore
	avg_coverage_area_pnode	média da área coberta pelas aproximações geométricas de uma árvore
	avg_overlap_area_pnode	média da área de sobreposição das aproximações geométricas de uma árvore
	avg_dead_space_pnode	média da área morta contida nas aproximações geométricas de uma árvore

Fonte: Autoria Própria

Parametrização dos índices. Esse grupo é composto por 4 atributos que indicam

Tabela 5 – Descrição das informações estatísticas sobre a execução de operações usando o índice.

Grupo	Coluna	Descrição
Informação estatística sobre execução de operações usando o índice	workload	tipo de workload realizado: Criação do índice, PQ, WQ, CQ
	table_name	Redundante em relação ao dataset_name
	index_type	tipo de índice: R-Tree, R*-Tree ou R Hilbert-Tree
	page_size	tamanho de página utilizado pelos nós
	buffer_type	tipo do buffer utilizado
	number_of_executions	número de execuções
	number_of_operations	número de operações executadas para cada tipo de workload
	avg_index_time	Tempo médio da operação
	avg_read_time	Tempo médio de leitura
	avg_write_time	Tempo médio de escrita
	avg_processed_entries_num	Média do número de entradas processadas
	avg_split_num	Média do número de splits realizados
	avg_visited_nodes_num	Média do número de nós visitados
	avg_written_nodes_num	Média do número de nós escritos
	stddev_index_time	Desvio padrão do tempo de indexação
	stddev_read_time	Devio padrão do tempo de leitura do índice para o conjunto de dados
	stddev_write_time	Devio padrão do tempo de escrita do índice para o conjunto de dados
	stddev_processed_entries_num	Desvio padrão do número de entradas processadas
	stddev_split_num	Devio padrão do número de splits realizados
	stddev_visited_nodes_num	Desvio padrão do número de nós visitados
	stddev_written_nodes_num	Desvio padrao do número de nós escritos

Fonte: Autoria Própria

a unicidade de um índice no dataset. Um índice é construído em um determinado dataset, usando um tamanho de página, uma técnica de buffer e um conjunto de parâmetros internos da árvore, como por exemplo a taxa de ocupação mínima dos nós.

Tabela 4 – Descrição da parametrização dos índices

Grupo	Coluna	Descrição
Configuração única de um índice	src_id	identificador do dataset. Redundante a coluna dataset_name
	bc_id	Identificador que faz referência ao tamanho de página. Redundante a coluna page_size
	sc_id	resume um conjunto de parâmetros internos da árvore. Ex: tipo de split, taxa de ocupação mínima dos nós.
	buf_id	tipo do buffer utilizado. Redundante em relação a coluna buffer_type

Fonte: Autoria Própria

Informações estatísticas sobre a execução de operações usando o índice. Após a execução dos workloads foi possível coletar uma série de medidas estatísticas relacionadas à execução das operações, como por exemplo o número médio de nós escritos e o tempo médio da operação de indexação. A Tabela 5 apresenta essas medidas coletadas e suas respectivas descrições.

7 CONCLUSÕES E TRABALHOS FUTUROS

Bancos de dados espaciais representam fenômenos na natureza e são empregados dos mais diversos modos, como por exemplo em serviços baseados em localização, agricultura de precisão, simulação de desastres naturais e defesa militar.

O volume de dados gerados por essas aplicações está em constante crescente, fazendo com esse tipo de informação seja recuperado da forma mais rápida possível. Objetos espaciais possuem uma estrutura complexa e necessitam de técnicas específicas para que isso seja possível, como por exemplo: técnicas de indexação espacial.

A escolha correta do índice espacial e suas configurações a serem utilizadas é uma tarefa difícil, porém fundamental para que a redução do custo computacional e consequentemente do tempo de espera seja otimizado.

Visando contribuir com área de pesquisa em estruturas de indexação espaciais, esse trabalho teve como objetivo realizar cargas de teste e coletar informações estatísticas de diferentes tipos de bases de dados indexadas com variações de algoritmos e configurações. Desse modo, o produto obtido dessa análise pode ser utilizado como um conjunto de dados de treinamento para um algoritmo de aprendizado de máquina capaz de prever o melhor índice e suas configurações para uma determinada base de dados.

Grande parte dos trabalhos presentes na literatura relacionados ao aprendizado de máquina focam nos algoritmos e métodos utilizados para resolver um problema específico e acabam deixando de lado uma etapa muito importante que é a processo de coleta dos dados utilizados.

A dificuldade deste trabalho está principalmente na realização de testes de desempenho em índices espaciais e seus diferentes tipos de configurações para que todas a informações necessárias sejam coletadas, sendo elas: informações sobre a estrutura e configuração dos índices e informações estatísticas coletadas à partir dos testes de desempenho. Além disso, características relacionadas à geometria dos dados também foram extraídos previamente.

Uma das limitações mais relevantes para o processo de extração desse conjunto de features é o custo computacional e tempo demorado para que os testes de desempenho sejam executados, o que inviabiliza a execução de bases maiores do que as utilizadas neste trabalho. Assim, seria necessário o uso de um hardware de alto poder computacional, tolerante a falhas de energia e totalmente dedicado para esta tarefa.

De forma a dar sequência ao objetivo deste trabalho, como uma primeira abordagem é pretendido desenvolver um método para previsão do tempo de execução de uma determinada consulta sobre uma base de dados espaciais. Esse método será fundamentado em um algoritmo de aprendizado de máquina treinado e testado a partir do conjunto de *features* desenvolvido neste presente trabalho.

Referências

BECKMANN, N. et al. The r*-tree: An efficient and robust access method for points and rectangles. In: **Proceedings of the 1990 ACM SIGMOD International Conference on Management of Data**. New York, NY, USA: Association for Computing Machinery, 1990. (SIGMOD '90), p. 322–331. ISBN 0897913655. Disponível em: <https://doi.org/10.1145/93597.98741>. Citado 2 vezes nas páginas 9 e 15.

BERTELLA, P. G. K. et al. The application of spatial approximations to spatial query processing: A systematic review of literature. In: **Brazilian Symposium on Databases**. [S.l.: s.n.], 2021. p. 229–240. Citado na página 14.

CARNIEL, A. C. Spatial information retrieval in digital ecosystems: A comprehensive survey. In: **International Conference on Management of Digital EcoSystems**. [S.l.: s.n.], 2020. p. 10–17. Citado 2 vezes nas páginas 9 e 13.

CARNIEL, A. C.; CIFERRI, R. R.; CIFERRI, C. D. Festival: A versatile framework for conducting experimental evaluations of spatial indices. **MethodsX**, v. 7, p. 100695, 2020. ISSN 2215-0161. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2215016119302717>. Citado 4 vezes nas páginas 9, 10, 11 e 17.

CARNIEL, A. C.; CIFERRI, R. R.; CIFERRI, C. D. A. The performance relation of spatial indexing on hard disk drives and solid state drives. In: **Brazilian Symposium on GeoInformatics**. [S.l.: s.n.], 2016. p. 263–274. Citado na página 9.

CARNIEL, A. C.; CIFERRI, R. R.; CIFERRI, C. D. A. Analyzing the performance of spatial indices on hard disk drives and flash-based solid state drives. **Journal of Information and Data Management**, v. 8, n. 1, p. 34–49, 2017. Citado na página 9.

CARNIEL, A. C.; SCHNEIDER, M. A survey of fuzzy approaches in spatial data science. In: **IEEE International Conference on Fuzzy Systems**. [S.l.: s.n.], 2021. p. 1–6. Citado na página 9.

EGENHOFER, M.; HERRING, J. Categorizing binary topological relations between regions, lines and points in geographic databases, the 9-intersection: Formalism and its use for naturallanguage spatial predicates. **Santa Barbara CA National Center for Geographic Information and Analysis Technical Report**, v. 94, p. 1–28, 01 1990. Citado na página 13.

GUTTMAN, A. R-trees: A dynamic index structure for spatial searching. In: **Proceedings of the 1984 ACM SIGMOD International Conference on Management of Data**. New York, NY, USA: Association for Computing Machinery, 1984. (SIGMOD '84), p. 47–57. ISBN 0897911288. Disponível em: <https://doi.org/10.1145/602259.602266>. Citado 2 vezes nas páginas 9 e 15.

KAMEL, I.; FALOUTSOS, C. Hilbert r-tree: An improved r-tree using fractals. In: **Proceedings of the 20th International Conference on Very Large Data Bases**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994. (VLDB '94), p. 500–509. ISBN 1558601538. Citado 2 vezes nas páginas 9 e 15.

KATIYAR, P. et al. Spiderweb: A spatial data generator on the web. In: **Proceedings of the 28th International Conference on Advances in Geographic Information Systems**. New York, NY, USA: Association for Computing Machinery, 2020. (SIGSPATIAL '20), p. 465–468. ISBN 9781450380195. Disponível em: <https://doi.org/10.1145/3397536.3422351>. Citado 2 vezes nas páginas 11 e 16.

MARCUS, R. et al. Neo: A learned query optimizer. **Proc. VLDB Endow.**, VLDB Endowment, v. 12, n. 11, p. 1705–1718, jul 2019. ISSN 2150-8097. Disponível em: <https://doi.org/10.14778/3342263.3342644>. Citado na página 10.

OpenStreetMap contributors. **Planet dump retrieved from https://planet.osm.org** . 2017. <https://www.openstreetmap.org>. Citado na página 11.

SELLIS, T. K.; ROUSSOPOULOS, N.; FALOUTSOS, C. The r+-tree: A dynamic index for multi-dimensional objects. In: **Proceedings of the 13th International Conference on Very Large Data Bases**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1987. (VLDB '87), p. 507–518. ISBN 093461346X. Citado na página 15.