

Agrupamento de Eventos de Trânsito baseado em Tesselção

Mariana M. Garcez Duarte^{1*}, Marcos V. Pontarolo^{1 †},
Rebeca Schroeder Freitas², Carmem S. Hara¹

¹Departamento de Informática, UFPR - Curitiba - PR

²Departamento de Ciências da Computação, UDESC - Joinville - SC
{mmgduarte,.mvp19, carmem}@inf.ufpr.br,
rebeca.schroeder@udesc.br

Abstract. *In the urban environment, data collected from traffic events can serve as elements of study for city planning. The challenge is to transform this raw data into knowledge of mobility. Because of the speed data is reported, each event is usually stored as an individual record. Although this approach guarantees fast insertions in a database, it produces poor performance for queries that retrieve events that satisfy spatio-temporal filters. To address this problem, this paper proposes a method for storing and indexing traffic events based on a tessellation of the area of interest. We demonstrate the proposed method through a case study on data collected by Waze.*

Resumo. *Dados coletados a partir de eventos no trânsito, como engarrafamentos e acidentes, são importantes para o planejamento da mobilidade em cidades. O desafio é transformar este conjunto de dados em conhecimento sobre mobilidade. Em razão da velocidade com que os dados são reportados, o armazenamento de eventos é geralmente feito como registros individuais. Embora este modelo de armazenamento garanta um baixo custo de inserção em uma base de dados, ele produz baixo desempenho em consultas que necessitam recuperar eventos que satisfaçam filtros espaço-temporais. Para tratar deste problema, este artigo propõe um método para o armazenamento e indexação de dados de eventos de trânsito baseado em uma tesselação da área de interesse. É apresentado um estudo de caso que demonstra sua aplicação sobre dados coletados no aplicativo Waze.*

1. Introdução

No ambiente urbano, dados coletados a partir de eventos no trânsito podem ser elementos de estudo para fomentar o planejamento de cidades. Engarrafamentos, acidentes e encontros são exemplos de eventos em um ambiente urbano que afetam a mobilidade. O desafio é transformar este conjunto de dados espaço-temporais em conhecimento sobre mobilidade. Em razão da velocidade com que os dados são reportados, o armazenamento é geralmente feito em registros individuais. Embora este modelo de armazenamento garanta um baixo custo de inserção em uma base de dados, ele produz baixo desempenho em consultas espaço-temporais. Isso se deve à necessidade de verificar, para cada registro, se os filtros da consulta são satisfeitos. [Khan and Kannapiran 2019] afirmam que,

*Bolsista Capes do Mestrado em Informática - PPGInf - UFPR

†Bolsista PET Computação-UFPR

em geral, os principais problemas na indexação e gerenciamento de dados espaciais são a alta demanda e o desempenho dos recursos de computação, bem como um maior tempo de resposta das consultas. Para tratar desse desafio, este artigo propõe um **Método de Indexação de Dados de Mobilidade (MIDM)**.

O método MIDM foi proposto para aplicações baseadas em armazenamento de eventos, que são caracterizadas pela existência apenas de operações de inserção de dados. O MIDM contempla três estratégias: indexação espacial, indexação temporal e associação de objetos à indexação espaço-temporal. Para a indexação espacial, a estratégia baseia-se na tesselação, ou seja, na divisão em grade da área geográfica de interesse, formando células geográficas (CGs), como ilustrado na Figura 1. Dessa forma, eventos que ocorrem na mesma CG são agrupados em blocos, garantindo assim que sejam armazenados de forma contínua no disco. Eventos possuem intrinsecamente uma dimensão temporal. Como o período de monitoramento dos eventos pode ser longo, o MIDM divide os blocos por períodos de tempo e cria um índice B+ sobre estes períodos. A estratégia de associação de objetos à indexação espaço-temporal é baseada em índices bitmap. Ruas, bairros e pontos de interesse, como parques e escolas, são exemplos de objetos no contexto urbano. Tais objetos podem ser considerados como chave de um índice bitmap, no qual o vetor de bits é indexado pelas células geográficas. Neste exemplo, a chave do índice são as ruas da cidade e os vetores de bits de cada rua têm o valor 1 nas células geográficas nas quais há um evento de engarramento. As estratégias propostas pelo MIDM são apresentadas em maior detalhe na Seção 2.

Como análise preliminar do método proposto, na Seção 3 é apresentado um estudo de caso de aplicação do MIDM sobre dados coletados pelo aplicativo Waze na cidade de Joinville-SC. A Seção 4 apresenta trabalhos relacionados e uma comparação com o método proposto. O artigo conclui na Seção 5 apresentando trabalhos futuros.

2. O Método MIDM

O método proposto neste artigo tem como objetivo reduzir o custo computacional do processamento de consultas de mobilidade. Seu diferencial é a combinação da utilização de bitmaps e a distribuição dos registros pela sua localidade geográfica e pela dimensão temporal. O método é baseado em 3 estratégias: (1) indexação espacial, (2) indexação temporal e (3) associação de objetos à indexação espaço-temporal.

2.1. Estratégia para Indexação Espacial

Para facilitar o processamento de consultas que envolvem a dimensão espacial, é aplicada a técnica de tesselação, na qual a área de interesse é dividida em uma grade, que forma células geográficas (CGs). Dessa forma, eventos do mesmo tipo que ocorrem dentro dos limites de uma CG são armazenados em conjunto, formando um *bloco*, podendo existir n blocos em um mesmo CG dependendo da quantidade de eventos que ocorreram na mesma área. Para ilustrar, considere o exemplo da Figura 1. Eventos são representados por quadrados nas CGs, e blocos tem um tamanho pré-determinado. Quando o bloco atinge seu tamanho limite, ele é transferido para o armazenamento em disco. Para obter todos os blocos que se referem à mesma CG, é criado um vetor de blocos. No vetor, cada elemento refere-se a uma CG e aponta para uma lista de blocos que contém eventos que ocorrem naquela célula.

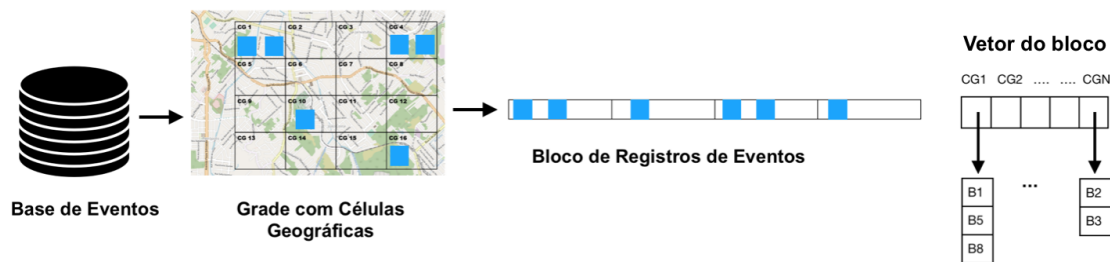


Figura 1. Indexação Espacial

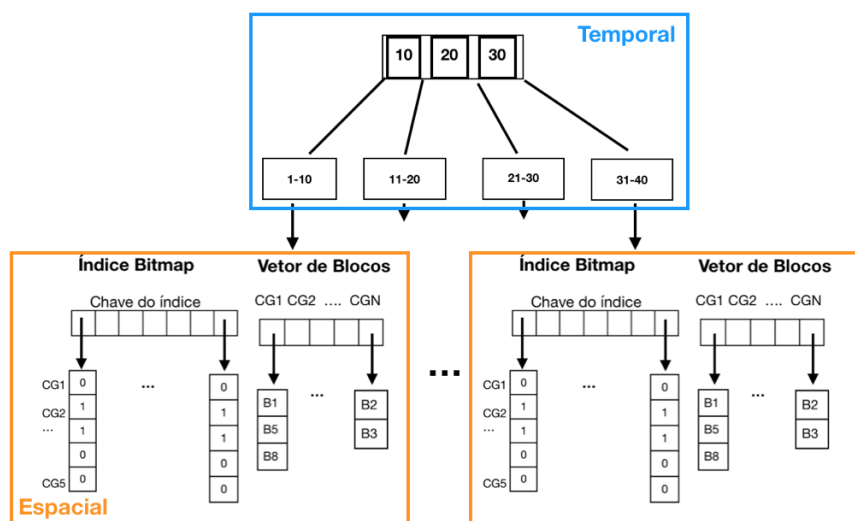


Figura 2. Associação de Objetos à Indexação Espaço-Temporal

2.2. Estratégia para Indexação Temporal

Cada registro contém a informação do instante em que o evento ocorreu. Para a criação dos blocos de registros, considera-se que eles são recebidos pelo sistema em ordem cronológica. Dessa forma, cada bloco está associado a um período de tempo, definido pelos *timestamps* do primeiro e último registro do bloco. Estes períodos, em conjunto com a lista de blocos associados a cada CG, podem ser utilizados para processar consultas com filtros espaço-temporais. No entanto, esta lista de blocos pode tornar-se excessivamente longa, o que pode gerar ineficiência para processar os filtros temporais. Para evitar este problema, o MIDM divide o monitoramento dos eventos em segmentos de tempo (que podem ser horas, dias, semanas, ou meses) e cria uma árvore B+ sobre estes segmentos, como ilustrado em "TEMPORAL"(Figura 2). Cada folha da árvore corresponde a um segmento de tempo e está associada a um vetor de blocos com eventos que ocorreram durante o segmento.

Dessa forma, o processamento de uma consulta espaço-temporal inicia percorrendo a árvore temporal para obter o(s) vetor(es) de blocos que possuem registros no período da consulta. A partir deste(s) vetor(es), os blocos das CGs que sobrepõem a área geográfica da consulta são recuperados para que os os registros que satisfazem aos filtros da consulta compõem o resultado.

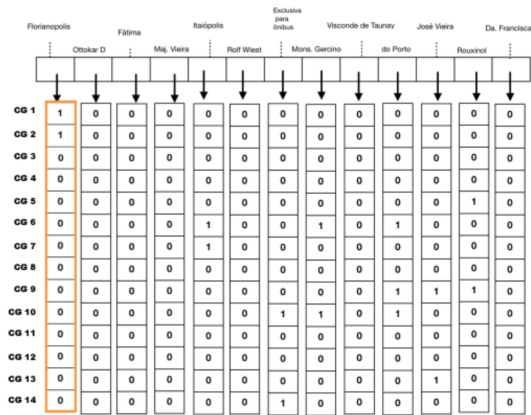


Figura 3. Índice bitmap sobre engarrafamentos com chave rua

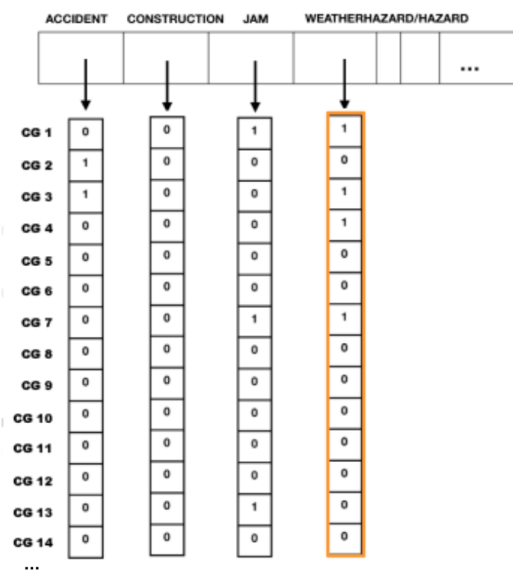


Figura 4. Índice bitmap sobre alertas com chave tipo de alerta

2.3. Estratégia para Associação de Objetos

O processo descrito na Seção 2.2 potencialmente reduz o espaço de busca de uma consulta para um pequeno número de CGs que contém a área de interesse. No entanto, é possível filtrar ainda mais as CGs que podem conter resultados da consulta se houver uma associação das CGs com objetos (como ruas e parques), bem como de classes (como tipo de ocorrência – colisão, atropelamento). Por exemplo, uma longa rua pode estar contida em mais de uma CG, mas existem colisões apenas em trechos da rua de uma única CG. Assim, apenas blocos desta CG precisam ser recuperados.

Para facilitar o processamento de consultas que envolvem objetos, o MIDM propõe a utilização de índices bitmap. Os valores da chave do índice consistem no domínio do objeto e os vetores de bit são indexados pelas CGs. A Figura 3 apresenta um exemplo sobre eventos de engarrafamento. A chave do índice é o nome da rua e cada vetor de bits $vBit[Rua X]$ contém o valor 1 nas CGs que contém trechos da Rua X que apresentaram algum engarrafamento. Por exemplo, a Rua Florianópolis contém engarrafamento nas CGs 1 e 2. A Figura 4 mostra o índice bitmap de tipo de alerta. A chave do índice é o tipo de alerta e cada vetor de bits $vBit[Tipo Alerta]$ contém o valor 1 nas CGs que contém alerta do tipo correspondente. Por exemplo, o alerta de clima (*Weather Hazard*) está presente nas CGs 1, 3, 4 e 7.

Os vetores de bits dos índices bitmap sempre são indexados sobre CGs para que seja possível utilizar as operações tradicionais de índice bitmap, como conjunção e disjunção. Como os índices são sobre CGs, que são separadas por períodos de tempo, os índices bitmap também são criados para cada período, como ilustrado na Figura 2. Assim, a associação de objetos à indexação espaço-temporal é feita por meio do armazenamento na *árvore temporal* dos vetores de blocos e de índices bitmaps.

O MIDM tem como objetivo dar suporte a aplicações que requerem o processamento de consultas espaço-temporais e por objetos. Na próxima seção, é apresentado um

estudo de caso que mostra sua utilização sobre a base de dados do aplicativo Waze.

3. Estudo de Caso

Esta seção apresenta um estudo de caso utilizando a base de dados do aplicativo Waze¹ da cidade de Joinville-SC. O tamanho da base é de 16 GB, contendo dados no período de setembro 2017 a setembro 2018.

Para exemplificar o MIDM, foram extraídas duas tabelas do banco de dados do Waze (Figura 5). Uma possui dados sobre engarrafamento e a outra sobre alertas. O campo *line* indica o *conjunto de pontos* do engarrafamento e o *ponto* do alerta. Ao final da leitura das tabelas, todos os registros são armazenados em blocos no arquivo. *Blocos* são unidades de transferência entre o disco e a memória e contêm registros de um mesmo tipo de evento e que ocorreram em uma mesma CG. A Figura 6 representa o arquivo de registros de engarrafamento resultante da distribuição dos registros em CGs e o seu armazenamento pelo MIDM. O mesmo processo foi realizado para os registros de alerta (Figura 7).

A consulta considerada é: "Quais são os trechos da rua Florianópolis que tiveram engarrafamento e alerta de clima em Janeiro de 2018?". A consulta é espaço-temporal e envolve as operações de seleção, junção e projeção. Ela foi escolhida para avaliar o impacto da utilização de bitmaps para a junção de diferentes tipos de eventos e a utilização de blocos. A consulta na base é apresentada a seguir. A consulta realiza a filtragem pelo período de tempo, pelo tipo de alerta (*Weather Hazard*), por engarrafamentos e alertas que ocorreram da Rua Florianópolis e por eventos que intersectam no espaço. Para executar esta consulta, seria necessário realizar a interseção (espacial) de todos os registros de alerta de clima da Rua Florianópolis e os de engarrafamento. A estratégia proposta pelo MIDM reduz a quantidade de registros considerados para a operação de interseção: são comparados apenas os registros que ocorrem nos mesmos CGs, como detalhado a seguir.

```
SELECT i.geom
FROM waze.jams i,waze.alerts j
WHERE i.pub_utc_date > '2017-12-31 23:59:59'
      and j.pub_utc_date > '2017-12-31 23:59:59'
      and i.pub_utc_date < '2018-01-08 00:00:00'
      and j.pub_utc_date < '2018-01-08 00:00:00'
      and j.type= "WeatherHazard"
      and j.name= "R. Florianópolis"
      and i.name=j.name
      and st_Intersects(i.geom,j.geom);
```

O processamento da consulta inicia com a seleção do período de Janeiro de 2018 na árvore temporal. Todas as semanas na subárvore (que correspondem às semanas do mês) são consideradas. Assim, o conjunto de vetores de bloco e os índices de bitmap correspondentes a estas semanas são selecionados. São identificados os objetos envolvidos na consulta para determinar se há índices bitmap que podem ser utilizados. Neste

¹<https://github.com/joinville/Joinville-Smart-Mobility>

as estratégias adotadas pelo MIDM e apresenta uma análise comparativa com base na consulta apresentada.

A árvore STIG [Doraiswamy et al. 2016], baseada em GPUs, corresponde a uma árvore Kd em seus nodos e suas folhas são compostas por partições, com o intuito de processar em paralelo consultas espaço-temporais sobre dados históricos. O aspecto temporal é apenas um atributo. Com a árvore STIG [Doraiswamy et al. 2016] a consulta é otimizada com GPU e duas árvores kd são utilizadas para os dois tipos de ocorrência, alerta e engarrafamento. As árvores kd tem dimensão espacial e temporal. Portanto o índice pode ser utilizado para obter os alertas com os filtros espaço-temporais a partir de uma árvore STIG e os engarrafamentos a partir de uma segunda árvore STIG. No entanto, para cada indexação por tipo de objeto seria criada uma nova dimensão, o que pode resultar em uma árvore com registros esparsos. A árvore STIG [Doraiswamy et al. 2016] orientou a criação de blocos de registros no método proposto.

A Árvore Buneman [Buneman et al. 2004], realiza o versionamento de uma base de dados. Esta estrutura é otimizada para buscas históricas. Um conjunto de experimentos demonstra que o arquivo não gera sobrecarga de espaço de armazenamento. Ao utilizar a árvore Buneman [Buneman et al. 2004] para a consulta do estudo de caso, poderiam ser selecionados os tipos de evento engarrafamento e alerta, e todos os registros correspondentes ao período seriam procurados necessitando fazer a combinação de eventos de engarrafamento e alerta na rua Florianópolis, pois a parte geográfica não é tratada. A árvore Buneman [Buneman et al. 2004] orientou a criação da árvore temporal e a indexação por objetos.

O índice HSB [Siqueira et al. 2012] é um índice espacial baseado em árvore com cada nodo folha apontando para um vetor de bits de um índice Bitmap de junção estrela. É utilizada a implementação de bitmap chamada FastBit. Consultas de baixa seletividade beneficiam-se com o índice HSB. Para responder à consulta apresentada na Seção 3, o índice HSB poderia ser utilizado para filtrar as regiões geográficas. Seriam utilizados dois índices HSB, um para cada tipo de evento, engarrafamento e alerta. Como não há um filtro temporal, não há pré-seleção que considere o período solicitado. Portanto, a busca envolveria todos os registros resultantes da conjunção independente do tempo. O índice HSB [Siqueira et al. 2012] orientou a utilização de índices bitmap.

O método proposto combina partes dos trabalhos relacionados para otimizar tanto aspectos temporais como geográficos para consultas de mobilidade e a utilização de bitmaps para diminuir o espaço de busca. A árvore STIG [Doraiswamy et al. 2016] orientou a criação de blocos de registros no método proposto. A árvore Buneman [Buneman et al. 2004] orientou a criação da árvore temporal e a indexação por objetos. O índice HSB [Siqueira et al. 2012] orientou a utilização de índices bitmap.

5. Considerações

Este artigo apresentou um método para a indexação de dados de mobilidade (MIDM). O diferencial da proposta é a combinação da utilização de bitmaps e a distribuição dos registros pela informação espaço-temporal. O método é baseado em 3 estratégias: (1) indexação espacial, (2) indexação temporal e (3) associação de objetos à indexação espaço-temporal. O método visa diminuir o tempo de busca e tornar o processamento de dados de mobilidade eficiente. O estudo de caso apresentado utilizou dados coletados pelo

aplicativo Waze. Atualmente, o método encontra-se em processo de implementação. Portanto, como trabalho futuro, o método será avaliado com experimentos utilizando dados reais para avaliar o desempenho em consultas, bem como o custo da indexação. Entretanto, vale ressaltar que, mesmo a aplicação tendo sido realizada sobre dados de mobilidade, a proposta de indexação de pontos no espaço pode ser aplicada em outros contextos.

A forma como o índice foi modelado pode comprometer a sua dinamicidade geográfica. A grade geográfica deve ser definida a priori, o que necessita que o usuário defina um tamanho adequado, definido pela quantidade de linhas e colunas e, desconsidere a densidade de entradas de acordo com a localização geográfica e as variações de densidade nas regiões em diferentes instantes de tempo. Para encontrar o tamanho ótimo para a grade geográfica com um dado conjunto de dados, ou seja, o melhor tamanho, é necessária a realização de uma análise experimental. Além disso, uma possível solução para o problema de densidade dos dados seria a utilização de divisões dentro de CGs, dividindo as que possuem alta densidade. Porém, essa abordagem aumentaria o tamanho da representação do MIDM, pois na modelagem atual, que possui as CGs de tamanho regular, a grade não é efetivamente criada, apenas calculada a partir do tamanho definido. Assim, é possível saber se um evento pertence a uma CG somente pelo cálculo. Caso fosse implementada a sub-divisão, seria necessário manter a informação de quais CGs foram divididas. Além disso, seria preciso modificar a criação dos bitmaps, podendo utilizar duas hierarquias de bitmap, um bitmap para as CGs completas e um bitmap para cada CG que tenha sido sub-dividida. Caso uma CG tenha sido dividida em um dado período de tempo, seria necessário que o bitmap criado para ela fosse criado para todos os períodos existentes para realizar a conjunção em consultas que utilizam diversos períodos.

Referências

- Buneman, P., Khanna, S., Tajima, K., and Wang-Chiew (2004). Archiving Scientific Data. *ACM Transactions on Database Systems (TODS)*, 29(1):2–42.
- Doraiswamy, H., Vo, H. T., Silva, C. T., and Freire, J. (2016). A gpu-based index to support interactive spatio-temporal queries over historical data. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pages 1086–1097.
- Khan, S. and Kannapiran, T. (2019). Indexing issues in spatial big data management. *SSRN Electronic Journal*.
- Mahmood, A. R., Punni, S., and Aref, W. G. (2019). Spatio-temporal access methods: a survey (2010 - 2017). *GeoInformatica*, 23(1):1–36.
- Siqueira, T. L. L., de Aguiar Ciferri, C. D., Times, V. C., and Ciferri, R. R. (2012). The SB-index and the HSB-index: efficient indices for spatial data warehouses. *Geoinformatica*, 16(1):165–205.