Novo Método de Enquadramento de Objetos Espaciais Complexos em Histogramas Espaciais

Isabella de Freitas Nunes[1], **Thiago Borges de Oliveira**[1]

[1]Instituto de Ciências Exatas e Tecnológicas (ICET)
Universidade Federal de Goiás, Regional Jataí (UFG)
BR 364, KM 195, 3800 – 75.801-615 – Jataí – GO – Brasil

idfn.ufg@gmail.com, thborges@ufg.br

Abstract. *The selectivity estimate is an important metric when selecting efficient execution plans on spatial databases. However, little effort was dedicated to enhance the methods and data structures which support the calculations of these estimations. In this paper we proposed an enhancement in the method used to make a multidimensional grid histogram. The proposed method reduced the error in the estimation up to 30.16%, when estimating the cardinality of spatial window queries, compared to the grid histogram construction method originally proposed.*

Resumo. *A estimativa de seletividade é uma importante métrica para escolha de planos de execução eficientes em banco de dados espaciais. No entanto, pouco estudo foi dedicado ao aprimoramento dos métodos e estruturas de dados que formam a base para o cálculo desta estimativa. Neste artigo propomos um aprimoramento no método de construção dos histogramas multidimensionais de grade, que resultou numa redução do erro de estimativa de até 30.16%, ao estimar a cardinalidade do conjunto resultante de consultas espaciais de janela, comparado com a técnica original de histograma de grade para dados espaciais.*

1. Introdução

Os dados espaciais, representam elementos do mundo real, descrevendo aspectos geográficos e espaciais de fenômenos da superfície terrestre (edificações, ruas, rios, áreas de vegetação, acidentes geográficos e outros) [Rigaux et al. 2002], sendo muitas vezes utilizados para auxiliar a tomada de decisões em larga escala e de grandes organizações. A fim de armazenar, recuperar, combinar, realizar análises variadas acerca de uma região, além de confeccionar materiais cartográficos e outros, estes dados são processados em SDBMS (*Spatial Database Management System*) [Campbell and Shin 2012].

Ao realizar uma consulta, um SDBMS deve ser capaz de definir qual o melhor plano de execução, a fim de conduzir a execução das consultas de forma eficiente. Esta escolha é realizada através de métricas registradas nos metadados do SDBMS. Uma métrica bastante empregada para este fim é a estimativa de seletividade. De acordo com [An et al. 2001], estimar a seletividade de consultas é crucial em um otimizador de consultas, a fim de que o melhor plano de execução seja escolhido.

Uma das técnicas propostas para estimar a seletividade de consultas espaciais é o uso do histograma multidimensional, uma estrutura de dados cuja principal característica

é a divisão da extensão espacial do *dataset* (base de dados) em *buckets* (células), que registram a quantidade de objetos espaciais no fragmento do espaço do *dataset* [Mamoulis and Papadias 2001]. Um tipo de histograma multidimensional é o histograma de grade, onde os *buckets* possuem tamanho fixo. Para [Acharya et al. 1999], o uso de histogramas tornou-se popular devido a sua construção ser simples, bem como a pouca utilização de espaço de armazenamento, além de não necessitar que a distribuição da entrada seja conhecida previamente.

Como o histograma é uma técnica de aproximação de um *dataset*, a geometria dos objetos espaciais (linhas, polígonos, e pontos) é também aproximada através de algumas estruturas, que mantém propriedades geométricas essenciais (estruturas conservativas) [Brinkhoff et al. 1994]. De acordo com [Gatti 2000], a estrutura mais utilizada é o mínimo retângulo envolvente (MBR - *minimum bounding rectangle*), ou seja, o menor retângulo com lados paralelos aos eixos das dimensões, que envolve todo o objeto espacial. Suas principais vantagens são o pouco espaço de armazenamento que ocupa e o baixo custo computacional da avaliação inicial dos predicados espaciais (etapa de filtragem).

No entanto, devido ser uma aproximação bem simples de um objeto espacial, o MBR causa problemas em estruturas de dados, devido aos erros de sua aproximação. O principal problema do MBR é a área morta (*dead space*), ou seja, o espaço livre que não é preenchido pelo objeto original. Para um *dataset* grande e composto de muitos objetos complexos, do tipo linha, a área morta pode interferir significativamente na construção das estruturas de dados. Como os histogramas de grade usam o MBR dos objetos, isso pode tornar a estimativa de seletividade das consultas espaciais bem divergente da realidade, e levar à escolha de um plano de execução ineficaz para uma consulta.

Alguns métodos alternativos para aproximar objetos espaciais foram propostos em [Brinkhoff and Kriegel 1994], [Brinkhoff et al. 1994], e [Lee et al. 1996], visando diminuir a área morta das aproximações. Este artigo é um resultado parcial de um projeto de pesquisa, cujo objetivo é investigar como os métodos de aproximação existentes interferem na estimativa de seletividade das consultas espaciais e escolher ou propor um novo método com o intuito de melhorar tal estimativa. Neste artigo, apresentamos o resultado de experimentos que comprovam que o uso de melhores aproximações dos objetos espaciais podem resultar numa melhora significativa da estimativa de seletividade.

O restante do texto está organizado da seguinte forma: na Seção 2 é apresentado um referencial teórico com conceitos relevantes para o entendimento do método proposto, a Seção 3 apresenta os trabalhos relacionados, a Seção 4 apresenta o método inicial proposto, que realiza o enquadramento dos objetos espaciais complexos, a Seção 5 apresenta os resultados iniciais dos experimentos e por fim, a Seção 6 apresenta a conclusão e trabalhos futuros.

2. Referencial Teórico

2.1. Histograma Multidimensional

Uma das técnicas utilizadas para estimar a seletividade de consultas espaciais é o histograma multidimensional, uma estrutura de dados utilizada para simplificar um *dataset* real, cuja principal característica é a divisão da extensão espacial deste *dataset* em uma

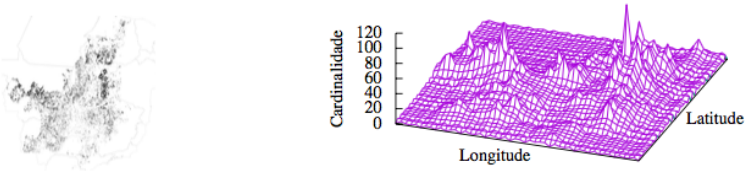

Figura 1. Dataset de alertas de desmatamento e histograma de grade [de Oliveira et al. 2015].

quantidade de *buckets*. De acordo com [Ioannidis and Poosala 1999], o histograma é uma técnica simples e de custo relativamente baixo, porém requer esforço computacional para calcular o número necessário de *buckets* e identificar os valores dos atributos que devem ser associados a eles, a fim de obter uma boa estimativa. A Figura 1 apresenta um dataset espacial de alertas de desmatamento do cerrado brasileiro e um gráfico em 3D, ilustrando os valores das cardinalidades para cada *bucket* em um histograma multidimensional de grade.

2.2. Aproximações de Objetos Espaciais

Os objetos espaciais também são aproximados no momento da construção do histograma. Trabalhar com a geometria real dos objetos é algo complexo, e por tal motivo é necessário utilizar representações simplificadas destes objetos espaciais [Gatti 2000]. Uma das técnicas mais utilizadas para representar objetos espaciais em um histograma é o MBR (*Minimum bounding rectangle* - mínimo retângulo envolvente). Tais aproximações conservam propriedades geométricas essenciais, tais como a posição no espaço e a extensão em cada dimensão [Teotônio 2008]. As duas maiores vantagens de se utilizar o MBR são o pouco espaço utilizado em seu armazenamento e a facilidade na avaliação de predicados espaciais.

3. Trabalhos Relacionados

O método proposto por [Mamoulis and Papadias 2001] enquadra os objetos espaciais na grade do histograma usando o centro do MBR, ignorando a extensão do objeto. Devido a isso, células do histograma, que contém o objeto originalmente, podem não registrar este fato, o que provoca erro na estimativa da cardinalidade do resultado das consultas.

De forma a aprimorar o método anterior, [de Oliveira et al. 2016] fez um estudo e propôs o método de sobreposição parcial, que calcula a sobreposição parcial do MBR em cada célula do histograma sobreposta, adicionando a fração obtida na contagem de objetos que é registrada em cada célula.

Apesar de proverem boas aproximações, conforme apresentado pelos autores, ambos os métodos possuem problemas com objetos do tipo linha, nos quais a extensão geográfica provoca o aumento do erro de seletividade. Um objeto do tipo linha naturalmente sobrepõe várias células do histograma e este fato deve ser registrado para se obter uma estimativa mais precisa. No entanto, o MBR de objetos do tipo linha sobrepõe muitas células do histograma erroneamente, devido a área morta. A próxima seção apresenta um método para lidar com estes dois problemas simultaneamente.

4. Método de Enquadramento de Objetos Espaciais Complexos

Nossa proposta baseia-se no método de sobreposição fracionada, proposto em [de Oliveira et al. 2016] e consiste no uso de uma aproximação mais refinada, para enquadrar melhor o objeto espacial nas células do histograma.

A Figura 2 ilustra esta operação. O método proposto consiste no seguinte procedimento: dado um objeto espacial do tipo linha, divide-se o mesmo em duas partes, gerando dois MBR's parciais. A área coberta pelos MBR's parciais é, frequentemente, menor que a área total do MBR original. Procura-se, então, o melhor ponto para dividir o objeto, de forma que a divisão minimize a área coberta pelos dois MBR's parciais (ou maximize a área morta, destacada na figura). O par de MBR's resultante é então usado para enquadrar o objeto no histograma. As células não sobrepostas pelos MBR's parciais não serão alteradas pelo enquadramento do objeto, e portanto, reduzirão o erro de estimativa sobre o histograma resultante.

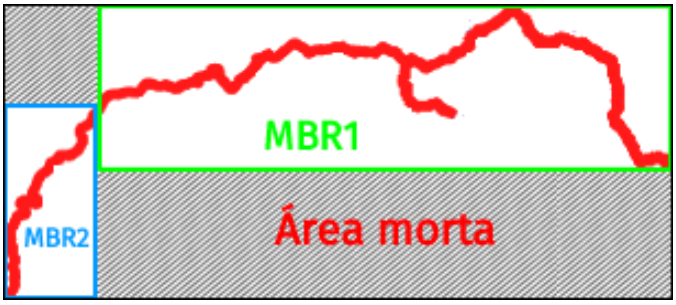

Figura 2. Ilustração do método de enquadramento proposto

5. Avaliação

Para realizar uma avaliação inicial do método proposto usou-se o *dataset* `ca_roads`, que contém 1.128.694 objetos espaciais do tipo linha, representando as ruas, avenidas e rodovias do estado da Califórnia - EUA. Este *dataset* é frequentemente empregado em experimentos com estruturas de dados espaciais.

O experimento executado consistiu em construir histogramas espaciais usando dois métodos de enquadramento de objetos: o método proposto em [Mamoulis and Papadias 2001], que usa o centro do MBR para enquadrar os objetos (mbrc), e a proposta descrita na Seção 4 (areafs). Para cada um dos histogramas construídos foi estimada[1] a cardinalidade do conjunto de resultados (c^e) para 500 consultas de janela distribuídas aleatoriamente e uniformemente pelo espaço geográfico do *dataset*. O resultado da estimativa (c^e) para cada consulta foi comparado com a cardinalidade real (c^r), obtida através da execução da consulta no *dataset*. Dado o conjunto de consultas $\mathcal{Q}$, o erro de estimativa η foi obtido somando o erro absoluto para cada consulta, conforme a Equação 1. Para avaliar a eficácia do método proposto, variou-se também as dimensões do histograma criado e o tamanho das consultas de janela, conforme indicado a seguir.

$$\eta = \sum_{q \in \mathcal{Q}} |c_q^r - c_q^e| \tag{1}$$

[1]A estimativa da cardinalidade do conjunto resultante de cada consulta foi obtida usando o método de estimativa sobre histogramas de grade proposto por [Mamoulis and Papadias 2001].

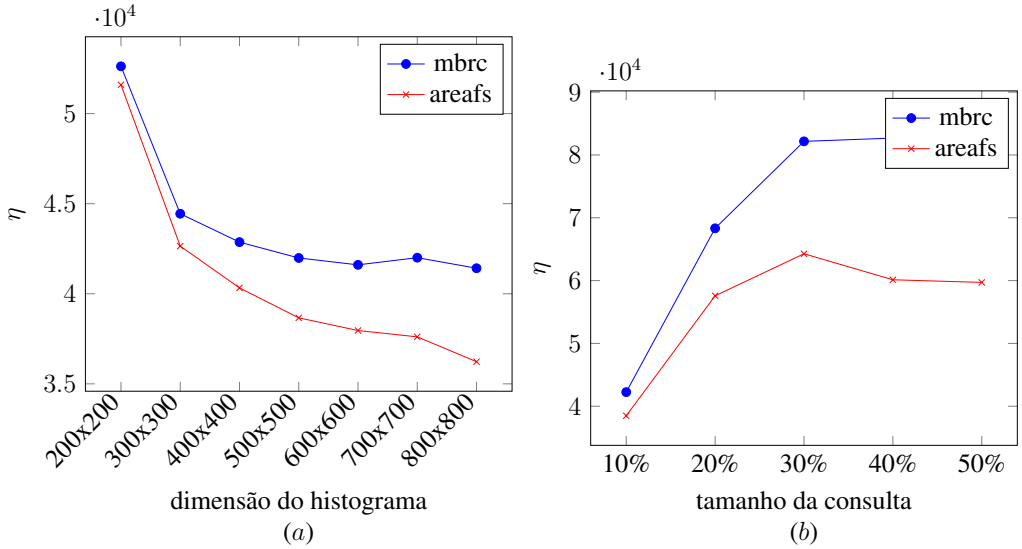

Figura 3. Comparação de η para os dois tipos de enquadramentos, variando as dimensões do histograma (a), e variando o tamanho da consulta (b).

O gráfico da Figura 5a apresenta o resultado do experimento com a variação da dimensão do histograma. Foram construídos histogramas de dimensões $200x200$, $300x300$, ..., $800x800$, usando os dois métodos de enquadramento. Observa-se que o erro diminui a medida que a dimensão do histograma aumenta. A proporção da redução também é maior nas maiores dimensões. No método (mbrc) o erro praticamente se estabiliza a partir do histograma com dimensão $500x500$. No método proposto (areafs), o erro continua diminuindo, mesmo para os histogramas com maior dimensão. A maior diferença de erro foi de 30.16%, para o histograma de dimensão $800x800$.

Avaliou-se também o efeito do tamanho das consultas na estimativa. Foram executadas consultas de tamanhos 10%, 20%, ..., 50%, calculados sobre o tamanho de cada dimensão do *dataset*, formando retângulos (janelas), posicionadas de forma aleatória e uniforme na área geográfica do *dataset*. Com as consultas de tamanho maior há uma tendência a aumentar o erro (valor de η), devido as mesmas retornarem mais objetos. O gráfico da Figura 5b apresenta os resultados obtidos. O método proposto (areafs) foi mais preciso que o método (mbrc) para todos os tamanhos de consultas testados.

6. Conclusão

A estimativa de seletividade é uma importante métrica para escolha de planos de execução eficientes em banco de dados espaciais. Neste artigo propusemos um aprimoramento no método de construção dos histogramas multidimensionais de grade, que resultou numa redução do erro de estimativa de até 30.16%, quando estimando a cardinalidade do conjunto resultante de consultas espaciais de janela, comparado com a técnica original de histograma de grade para dados espaciais.

O método proposto e experimentado procura exaustivamente a divisão ótima do MBR do objeto espacial, a qual retorna a menor área morta. Devido sua complexidade, pode não ser recomendado na prática para um banco de dados espacial. Na continuação

deste projeto, pretendemos investigar algoritmos mais eficientes para encontrar a divisão ótima ou uma boa divisão do MBR através de métodos heurísticos, bem como testar outras técnicas de aproximação dos objetos espaciais, como as propostas por [Lee et al. 1996]. Pretendemos ainda expandir o conjunto de experimentos, incluindo mais *datasets* de objetos espaciais complexos para confirmar a aplicabilidade do método.

Referências

Acharya, S., Poosala, V., and Ramaswamy, S. (1999). Selectivity estimation in spatial databases. *SIGMOD Rec.*, 28(2):13–24.

An, N., Yang, Z.-Y., and Sivasubramaniam, A. (2001). Selectivity estimation for spatial joins. In *Proceedings 17th International Conference on Data Engineering*, pages 368–375. IEEE.

Brinkhoff, T. and Kriegel, H.-P. (1994). *Approximations for a Multi-step Processing of Spatial Joins*, pages 25–34. Springer Berlin Heidelberg, Berlin, Heidelberg.

Brinkhoff, T., Kriegel, H.-P., Schneider, R., and Seeger, B. (1994). Multi-step processing of spatial joins. *SIGMOD Rec.*, 23(2):197–208.

Campbell, J. E. and Shin, M. (2012). *Geographic Information System Basics*.

de Oliveira, T. B., Costa, F. M., and Rodrigues, V. J. d. S. (2015). Definição de Planos de Execução Distribuídos para Consultas de Junção Espacial usando Histogramas Multidimensionais. In *Proceedings of the Brazilian Symposium on Databases*, pages 89–100, Petrópolis, RJ, Brazil.

de Oliveira, T. B., Costa, F. M., and Rodrigues, V. J. d. S. (2016). Distributed execution plans for multiway spatial join queries using multidimensional histograms. 7(1):To appear.

Gatti, S. D. (2000). Fatores que afetam o desempenho de métodos de junções espaciais: um estudo baseado em dados reais. Mestrado, UNICAMP, Campinas.

Ioannidis, Y. E. and Poosala, V. (1999). Histogram-based approximation of set-valued query-answers. In *Proceedings of the 25th International Conference on Very Large Data Bases*, VLDB '99, pages 174–185, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Lee, Y.-J., Lee, D.-M., Ryu, S.-J., and Chung, C.-W. (1996). *Controlled decomposition strategy for complex spatial objects*, pages 207–223. Springer Berlin Heidelberg, Berlin, Heidelberg.

Mamoulis, N. and Papadias, D. (2001). Multiway Spatial Joins. *ACM Transactions on Database Systems*, 26(4):424–475.

Rigaux, P., Scholl, M., and Voisard, A. (2002). *Spatial Databases: With Application to GIS*. Series in Data Management Systems. Morgan Kaufmann Publishers.

Teotônio, F. A. B. (2008). Comparação do desempenho dos ídices r-tree, grades fixas, e curvas de hilbert para consultas espaciais em bancos de dados geográficos. Mestrado do curso de pós-graduação em computação aplicada, Instituto Nacional de Pesquisas Espaciais - INPE, São José dos Campos.