poster:08

Estatística Zonal de Imagens de Sensoriamento Remoto Armazenadas em Banco de Dados NoSQL*

Adeline Marinho Maciel, Lúbia Vinhas, Gilberto Câmara

[1]Instituto Nacional de Pesquisas Espaciais (INPE)
Caixa Postal 515 – 12227-010 – São José dos Campos, SP – Brasil

adeline@dpi.inpe.br, {lubia.vinhas,gilberto.camara}@inpe.br

Abstract. *This paper shows how to implement zonal operations from Tomlin's Map Algebra in a multidimensional matricial database management system. The database consist of a set of remote sensing images stored in a SciDB instance. The zonal operator was used to compute spatial and temporal statistics over discrete regions of a vector map.*

Resumo. *Este artigo mostra como implementar operações zonais da álgebra de mapas de Tomlin em um sistema gerenciador de banco de dados matricial multidimensional. O banco de dados consiste de um conjunto de imagens de sensoriamento remoto armazenadas em uma instância do SciDB. O operador zonal foi usado para computar estatísticas espacial e temporal sobre regiões discretas de um mapa vetorial.*

1. Introdução

Dados de sensoriamento remoto são utilizados em diferentes pesquisas que tratam desde a detecção e monitoramento de áreas de desmatamento até a verificação de áreas propensas a riscos ambientais. Em geral são usados em conjunto outros dados geográficos, como dados hidrológicos, dados climáticos ou mapas geopolíticos. Por isso, a pesquisa em geoinformática sempre está em busca de algoritmos, tecnologias e ferramentas que permitam o tratamento de dados geográficos em diferentes representações de maneira eficiente e inovadora.

A tecnologia de bancos de dados direcionou uma evolução de arquitetura das aplicações geográficas, passando do armazenamento apenas da componente não espacial do dado até o gerenciamento de todo o dado geográfico através das extensões espaciais. Até o final dos anos 2000 esse desenvolvimento foi baseado somente nos modelos relacionais e objeto-relacionais para bancos de dados.

Com a disponibilidade de novos sensores de observação da Terra, experimentada nas últimas décadas, tem se pesquisado novas tecnologias para sistemas gerenciadores de banco de dados que sejam mais adequadas para o tratamento desses dados. Os bancos de dados matriciais multidimensionais tem recebido muita atenção como uma alternativa eficiente ao modelo relacional e objeto-relacional para bancos de dados. Além das capacidades de armazenamento é necessário investigar também como executar processamentos sobre esses dados de maneira eficiente, em particular usando ao máximo capacidades

*Os autores agradecem a CAPES e ao programa e-science da FAPESP (projeto temático 2014/08398-6) pelo apoio financeiro.

de processamento no lado do servidor e evitando assim a transferência de dados para o cliente.

Para a construção de aplicações em geoinformática e sensoriamento remoto é necessário muitos processamentos que precisariam estar disponíveis no lado do servidor. Por isso esse trabalho começa por considerar os processamentos que possam ser expressos através de conceitos da chamada *álgebra de mapas* [Tomlin 1990]. A álgebra de mapas de Tomlin (1990) é uma das formas tradicionais de manipulação de dados geográficos tanto para geração de novos dados quanto para a execução de análises para tomada de decisão.

Esse trabalho tem por objetivo explorar a implementação de operadores da álgebra de mapas em um ambiente de banco de dados matriciais multidimensionais, onde uma das dimensões é a dimensão temporal. Ou seja, onde os atributos espaciais variam no tempo. Um exemplo de dado espaço-temporal é uma série de imagens de sensoriamento remoto que são sistematicamente obtidas para uma dada região geográfica. Foram exploradas as operações zonais de Tomlin (1990) para extrair estatísticas zonais temporais de dados de sensoriamento remoto, mais especificamente produtos do sensor MODIS[1]. As restrições zonais utilizadas são dadas por polígonos de um mapa de desmatamento.

2. Álgebra de Mapas

O termo álgebra de mapas foi definido por Tomlin (1990) para indicar o conjunto de procedimentos de análise espacial em geoprocessamento que produz novos dados a partir de funções de manipulação aplicadas a um ou mais mapas. Os mapas são considerados como variáveis individuais e as operações definidas sobre elas são aplicadas de forma homogênea a todos os pontos do mapa [Barbosa et al. 1998]. Estas operações podem ser divididas em três tipos, de acordo com diferentes restrições espaciais:

- *Pontual* - O valor de um ponto, ou célula, no mapa de saída é determinado a partir de um único valor em um ou mais mapas de entrada, com localização correspondente.
- *Vizinhança ou Focal* - O valor da célula no mapa de saída é calculado a partir de uma vizinhança específica em torno da célula do mapa de entrada considerado.
- *Zonal* - O valor de cada célula no mapa de saída é determinado por todas as células contidos em uma mesma região, ou zona, do mapa de entrada, onde as restrições são fornecidas por outro mapa de entrada.

Diversos trabalhos utilizam a álgebra de mapas formalizada por Tomlin (1990) para manipulação de dados espaciais estáticos no tempo, mas poucos exploraram seu uso utilizando dados espaço-temporais. Dentre estes podemos destacar: Mennis et al. (2005), que propõem uma extensão da álgebra de mapas original para dados espaço-temporais, por meio da criação de novas funções; Frank (2005), onde o autor discute como a álgebra de mapas pode ser estendida e aplicada uniformemente para dados espaciais, temporais e espaço-temporal; e Mennis (2010), que apresenta uma extensão da álgebra de mapas convencional para tratar dados com até quatro dimensões, uma dimensão no tempo e três no espaço, denominada álgebra de mapas multidimensional.

[1]MODIS - (*Moderate Resolution Imaging Spectroradiometer*) sensor presente nos satélites TERRA e AQUA

Segundo Tomlin (1990) algumas das operações zonais mais comuns são: maioria zonal, máximo zonal, mínimo zonal, média zonal, soma zonal e estatística zonal. Diferente das outras operações que produzem como saída um novo mapa, a operação de estatística zonal toma dois mapas de entrada, um que contém os atributos que se deseja obter a estatística e outro com um conjunto de regiões, ou zonas, que são os elementos de análise. A saída é um dado tabular onde para cada zona, são calculadas um conjunto de estatísticas (por exemplo, valores máximo, mínimo, média, desvio padrão e variância) sobre os atributos dos elementos que estão espacialmente nela contidos.

3. Banco de dados

Devido à demanda por gerenciamento mais eficaz e a necessidade de se realizar análise sobre grandes volumes de dados, surgiram novas tecnologias de bancos de dados denominadas *NoSQL* (*Not only SQL*), que não requerem a criação de esquemas rígidos de armazenamento de dados e, na maioria, não utilizam a SQL como linguagem de consulta [Queiroz et al. 2013]. Uma dessas alternativas são os Bancos de Dados Matriciais Multidimensionais. Considerando que dados de observação da Terra, são capturados e disseminados em uma representação matricial com duas dimensões espaciais (na direção latitudinal e longitudinal). Além dessas, existe também a dimensão do atributo medido, por exemplo, um dado sensor produz imagens com várias bandas. Finalmente, a maioria das aplicações com esses dados requer que sejam observadas sequências de imagens ao longo do tempo. Por isso o interesse em explorar o uso dessa tecnologia de bancos de dados com conjuntos de dados de observação da Terra.

3.1. SciDB

O SciDB, *Scientific Database*, é um sistema de banco de dados matricial multidimensional desenvolvido a partir de um consórcio formado por diversos pesquisadores da área de banco de dados e representantes de comunidades científicas, como Sensoriamento Remoto e Astrofísica [Brown 2010]. O sistema possui duas linguagens de consulta para gerenciamento e análise de dados, a *Array Query Language* (AQL) e *Array Functional Language* (AFL) [Stonebraker 2012, Paradigm4, Inc. 2013].

O SciDB é projetado para trabalhar de maneira ótima em *clusters* de processadores, pois ele baseia-se no conceito de instâncias que rodam em nós diferentes, e são controlados por um nó gerenciador. Para o armazenamento o sistema pressupõe, o particionamento nas matrizes, nas suas diferentes dimensões, em *chunks* que são distribuídos nas diferentes instâncias rodando nos nós. Cada instância controla seus *chunks* de maneira independente. O nó coordenador é responsável pela execução das consultas, ou seja, orquestra a comunicação entre as instâncias disparando os processamentos que podem ser realizados nos nós locais e pela comunicação entre o *cluster* e as aplicações clientes.

Um ponto a ser notado é que o SciDB não tem nenhuma semântica associada ao dado. Não se pode dizer a que uma dada dimensão se refere (temporal, espacial ou radiométrica no exemplo abordado nesse trabalho), e por isso não oferece em sua linguagem de consulta operadores específicos para um tipo de dado (como os operadores de uma extensão espacial para banco de dados geográficos). Todas as operações oferecidas são aquelas baseadas em álgebra de matrizes.

4. Álgebra de Mapas em SciDB

Na álgebra de mapas, as operações zonais são aplicadas sobre um mapa de entrada e um mapa de restrição composto por conjunto de regiões, ou zonas, delimitadas por polígonos. Por se tratar de um dado geralmente discreto é bastante comum que tenham uma representação vetorial, ou seja, sejam representadas por geometrias como polígonos. O que gera uma dificuldade, pois não é possível, no SciDB, combinar dados com representação matricial e vetorial nem em termos de armazenamento, nem em termos de consulta.

Nesse trabalho a abordagem considerada é produzir, por pré-processamento, uma representação matricial para o mapa que contém as zonas e considerá-lo como mais uma matriz dentro do banco de dados. A partir daí foi possível a elaboração de consultas utilizando operadores do sistema SciDB como parte principal do processo de extração das estatísticas zonais temporais. Para isso, uma metodologia foi desenvolvida composta pelas seguintes etapas:

1. Seleção da região de estudo;
2. Pré-processamento e estruturação das imagens para armazenamento no SciDB;
3. Inserção do conjunto de imagens temporais e dos polígonos de desmatamento rasterizados, resultantes da etapa (2), como matrizes no banco de dados, por meio de operadores fornecidos pelo sistema;
4. Produção de consultas, com arranjo de operadores das linguagens AQL e AFL, para extração das estatísticas zonais da matriz multidimensional contendo as imagens, mapa de entrada, delimitada pela matriz contendo dados de desmatamento, mapa de restrição;
5. Geração de dado tabular contendo as estatísticas zonais para cada imagem, por meio do resultado obtido na etapa (4).

4.1. Exemplo

Para demonstrar o uso de operadores zonais em dados espaciais temporais armazenados no sistema, escolheu-se uma aplicação simples: a geração de estatísticas zonais de uma série temporal de um ano de produtos derivados de imagens do sensor MODIS a bordo do satélite TERRA. Em particular foi usado o índice EVI[2]. As restrições espaciais, ou zonas, foram os polígonos de desmatamento na mesma área das imagens. Ainda que esse seja apenas um exemplo ele é motivado pelo interesse em buscar métodos de detecção automática de perturbação na floresta, estudos em agricultura, e uso e cobertura da Terra. Executar tais análises em uma escala global e por longos períodos de tempo, é algo computacionalmente extensivo e por isso o interesse em estudar novos meios de armazenamento e processamento de grandes bases de dados de observação da Terra.

Os produtos MODIS foram obtidos de um banco de dados *online* da Embrapa - Empresa Brasileira de Agropecuária. Eles tem uma resolução de 250 metros e usou-se as imagens que cobrem o estado de Rondônia. Durante o ano de 2014, foram utilizadas 23 imagens, uma vez que estas imagens são produzidas a cada 16 dias; os polígonos de desmatamento utilizados são relativos ao município de Vale do Anari-RO, adquiridos do Projeto PRODES[3].

[2]EVI - *Enhanced Vegetation Index*

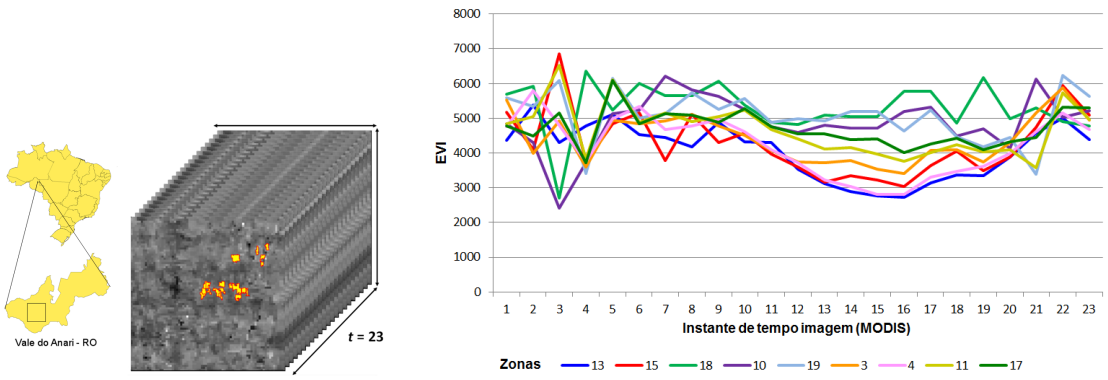
[3]`http://www.dpi.inpe.br/prodesdigital/`

A Figura 1 exibe um exemplo de consulta realizada no SciDB usando a linguagem AFL para extração das estatísticas zonais. Nesta consulta é feita a extração dos valores estatísticos da imagem de tempo 1, dimensão denominada *time_id*, do *array Input_MODIS* contendo 23 imagens, apenas para a zonas de restrição com o identificador 15.

```
aggregate(
    filter(
        project(
            cross_join(unpack(filter(Input_MODIS, time_id=1), y)
    as entr,unpack(Rest_Desm, x) as rest, entr.y,rest.x),
        entr.vals,rest.vals),
    rest.vals=15),
count(entr.vals) as length, sum(entr.vals) as summation, max(
   entr.vals) as maximum, min(entr.vals) as minimum, avg(entr.
   vals) as average, stdev(entr.vals) as staDevi, var(entr.vals)
    as variance)
```

Figura 1. Consulta para extração da estatística zonal de um instante de tempo e uma zona de restrição específicos

Como resultado desse processo foi gerado um dado tabular contendo todas as estatísticas zonais com valores de soma, máximo, mínimo, média, desvio padrão e variância para cada região de desmatamento em relação as imagens analisadas. A Figura 2(a) exibe a região de estudo e os polígonos de desmatamento de restrição sobre uma matriz contendo as 23 imagens armazenadas no banco de dados, numa estrutura de cubo. O gráfico da Figura 2(b) exibe a média dos valores de EVI extraídos para cada região de desmatamento a partir das 23 imagens.



(a) Estrutura de armazenamento (b) Média de EVI para cada região de desmatamento

Figura 2. Dados de estudo e série temporal para cada região explorada

5. Considerações Finais

Esse trabalho estudou como implementar um dos operadores zonais da álgebra de mapas de Tomlin em um sistema de banco de dados matricial multidimensional no gerenciador SciDB. Utilizou-se as linguagens fornecida pelo sistema para obter os resultados do operador.

O estudo mostrou que o modelo matricial para bancos de dados é bastante adequado para tratar grandes volumes de dados de observação da Terra, pois em geral, esses dados são obtidos em uma representação matricial (imagens). A multidimensional permite facilmente organizar também as múltiplas dimensões espectrais das imagens. E finalmente atende também a necessidade de se organizar as séries temporais de dados de imagens. A possibilidade de particionamento nas diferentes dimensões atende a necessidade de se trabalhar com "cubo" de dados que pode ser expandido em todas as dimensões.

O estudo mostrou também que, na aplicação pretendida, existe uma necessidade de se transformar dados com qualquer outra representação na representação matricial. Isso foi feito por pré-processamentos para harmonizar os diferentes matrizes em termos das dimensões espaciais. Por exemplo, a resolução espacial de todas as matrizes tem que ser a mesma. Apesar do custo de pré-processamento para transformação de representação, uma vez montado o banco, uma única consulta permitiu processar todos as células das matrizes, ficando o controle de particionamento da matriz entre as instâncias do SciDB por conta do sistema.

A continuidade desse trabalho será construir novos experimentos com volumes maiores de dados, a implementação de novos operadores, e testes mais robustos de eficiência.

Referências

Barbosa, C. C., Câmara, G., Medeiros, J. S., Crepani, E., Novo, E. M. L. M., and Cordeiro, J. P. C. (1998). Operadores zonais em álgebra de mapas e sua aplicação a zoneamento ecológico-econômico. In *Anais...*, pages 487–500, São José dos Campos. Simpósio Brasileiro de Sensoriamento Remoto, 9. (SBSR)., INPE.

Brown, P. G. (2010). Overview of SciDB: Large Scale Array Storage, Processing and Analysis. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, SIGMOD '10, pages 963–968, New York, NY, USA. ACM.

Frank, A. (2005). Map Algebra Extended with Functors for Temporal Data. In *Perspectives in Conceptual Modeling*, volume 3770 of *Lecture Notes in Computer Science*, pages 194–207. Springer Berlin Heidelberg.

Mennis, J. (2010). Multidimensional Map Algebra: Design and Implementation of a Spatio-Temporal GIS Processing Language. *Transactions in GIS*, 14(1):1–21.

Mennis, J., Viger, R., and Tomlin, C. D. (2005). Cubic Map Algebra Functions for Spatio-Temporal Analysis. *Cartography and Geographic Information Science*, 32(1):17–32.

Paradigm4, Inc. (2013). SciDB Reference Manual.

Queiroz, G. R., Monteiro, A. M. V., and Câmara, G. (2013). Bancos de dados geográficos e sistemas NoSQL: onde estamos e para onde vamos? *Revista Brasileira de Cartografia*, 3(65):479–492.

Stonebraker, M. (2012). SciDB: An Open-Source DBMS for Scientific Data. *ERCIM News*, (89).

Tomlin, C. D. (1990). *Geographic Information Systems and Cartographic Modeling*. Prentice-Hall Inc., New Jersey.