Automação do Pré-processamento de Séries Temporais de Dados da NOAA para Mineração de Dados

Wanderson Gomes de Almeida[2], Marcos Eduardo Gomes Borges[2], Pettras Leonardo Bueno dos Santos[2], Juliana Aparecida Anochi[2], Walter Abrahão dos Santos[1]

[1]Laboratório Associado de Computação e Matemática Aplicada – LAC
[2]Programa de Mestrado e Doutorado em Computação Aplicada – CAP
Instituto Nacional de Pesquisas Espaciais – INPE

{wandersonwmpp, marcoseborges, pettrasleonardo, juliana.anochi}@gmail.com, walter.abrahao@lac.inpe.br

Resumo. A complexa tarefa do estudo e previsão climática tem uma importância fundamental para as atividades humanas, devido aos efeitos causados pela dinâmica da atmosfera. Este artigo propõe um pré-processamento automatizado de dados climáticos de séries espaciais e temporais para posteriormente serem minerados por algum aplicativo de Mineração de Dados (DM). A abordagem adotada utiliza apenas soluções de código aberto e protótipos básicos de consultas SQL de estatísticas espaço-temporais de conjuntos de dados geo-referenciados em séries temporais disponíveis na *web* a partir do Centro Nacional de Pesquisa Atmosférica da agência de Administração Nacional do Oceano e da Atmosfera (NOAA), contribuindo para aumento de produtividade na análise de dados.

Palavras-chave. Sistemas de Informações Geográficas, Meteorologia, Mineração de Dados, Séries Temporais, NetCDF.

Abstract. The complex task of climate study and forecasting has a key importance to human activities from the effects caused by the atmosphere dynamics. This article proposes an automated pre-processing of spatial and time series climate data later to be mined by any application of Data Mining (DM). The approach taken adopts only open-source solutions and prototypes basic time-spatial statistical SQL queries from geo-referenced time series datasets available over the web from the National Center for Atmospheric Research of the National Oceanic and Atmospheric Administration (NOAA) catering to higher productivity on data analysis..

Keywords. Geographic Information Systems, Meteorology, Data Mining, Time Series, NetCDF.

1. Introdução

Devido aos efeitos da dinâmica da atmosfera, estudos e previsões climáticas são essenciais para permitir, por exemplo, prever se o próximo inverno será mais frio que a média, ou ainda, se haverá mais chuva que a estação anterior. O objetivo deste estudo é melhorar a análise dos dados de produtividade da fase de pré-processamento do processo de Descoberta de Conhecimento em Banco de Dados, do inglês *Knowledge Discovery in Databases* (KDD), através da avaliação das propriedades estatísticas espaço-temporais dos dados climáticos.

O processo de KDD é a descoberta da informação implícita e útil em grandes bases de dados e em geral emprega técnicas de alto nível de DM [Botia et al., 2002]. O pré-processamento é uma de suas três grandes fases, a qual é responsável por realizar as operações básicas de análise dos dados para definir, por exemplo, as estruturas das tabelas, os valores de potenciais para os atributos, formatos, tipos de dados e outras configurações.

Como estudo de caso, um sistema foi desenvolvido em Java como protótipo para extração dos dados empacotados no formato de arquivo netCDF, do inglês *Network Common Data Form*, da NOAA obtidos via FTP [NOAA, 2011]. Estes dados poderão ser exportados para um banco de dados PostGIS permitindo a coleta de estatísticas e a descoberta de conhecimento para os estudos da meteorologia e do clima.

Este artigo é organizado da seguinte forma. A seção 2 apresenta o formato de arquivo netCDF. A seção 3 focaliza na metodologia proposta. A seção 4 descreve um estudo de caso relatando o resultado encontrado. As conclusões são apresentadas na Seção 5 finalizando este trabalho.

2. O Formato de Dados NetCDF

Os arquivos de análises e previsões disponibilizados pela NOAA tem sido uma importante fonte de dados para os pesquisadores. Por isso, conhecer a estrutura e como lidar com o formato netCDF corretamente tem sido de grande importância devido ao seu uso cada vez mais constante. No entanto, as análises operacionais são afetadas por mudanças nos modelos, técnicas de análise, assimilação e uso de observações (ALMEIDA; VINHAS; CORREA, 1998). Por este motivo, surge a ideia de produzir uma consistente reanálise dos dados atmosféricos.

O formato de dados netCDF é um conjunto de interfaces com funções de acesso a dados armazenados na forma de matrizes, denominadas estruturas multidimensionais, para representação de dados científicos. Para manipular este aquivo de dados é preciso instalar uma biblioteca de plataforma independente desenvolvido por um grupo de pesquisadores do *Unidata Program Center*, em Boulder, Colorado [ESRL, 2011].

Um arquivo netCDF contendo dados que definem informações para uma aplicação particular é denominado “conjunto de dados”, do inglês *dataset*. Para um melhor armazenamento e agrupamento dos dados o arquivo netCDF pode ser organizado em: dimensões, variáveis e atributos. Um nome e um número de identificação pode ser atribuído aos arquivos para tornar possível a identificação das relações e atribuir um significado aos campos de dados existentes no *dataset* [UNIDATA, 2011].

3. Metodologia e Automação da Fase de Pré-processamento

Os arquivos de dados geográficos netCDF trazem medidas referentes ao período de janeiro de 1948 à dezembro de 2010 com uma resolução espacial de 2.5 graus em ambas as dimensões (latidute e longitude). Além disso, tem uma resolução temporal de 1 mês. As datas são codificadas com o número de horas (ou dias) a partir de 1-1-1 00:00:0.0 (ou 00:00:0.0 1800/01/01) em diante, sendo necessário a conversão do formato para leitura e compreensão humana.

Neste trabalho optou-se por utilizar o PostGIS, extensão espacial do SGBD PostgreSQL, destinada a trabalhar com dados geo-referenciados e multidimensionais. O PostGIS é um *plug-in* de código aberto robusto e compatível com os padrões definidos pelo Consórcio Geoespacial Aberto, do inglês *Open Geospatial Consortium* (OCG), que visa promover o desenvolvimento de padrões que facilitem a interoperabilidade entre Sistemas de Informações Geográficas, do inglês *Geographic Information Systems* (GIS).

Neste trabalho, foi decidido criar um protótipo de um aplicativo Java e ilustrar sua aplicação em apenas uma categoria simples de automação do pré-processamento dos dados fazendo uma análise espacial em tempo estatístico básico. Este aplicativo permite: (1) Abrir arquivos netCDF fornecidos pela NOAA, (2) Extrair as informações contidas no dataset (3) Conectá-los e armazená-los no banco de dados PostGIS e (4) Gerar estatísticas através de sentenças espaço-temporais. As tabelas para o armazenamento das informações são criadas automaticamente pelo *software*, o que facilita a manipulação das informações.

4. Estudo de Caso e seus Resultados

Como todas as informações fornecidas pela NOAA são referenciadas pela latitude e longitude, a geoinformática desempenha um papel fundamental, uma vez que adiciona facilidades de processamento espacial para a solução proposta. Para executar as estatísticas, as consultas no banco de dados geográficos foram feitas a fim de selecionar as medidas localizadas no estado de São Paulo. As informações geográficas vetoriais com a localização dos municípios que compõe o estado de São Paulo foram baixadas do site do IBGE [IBGE, 2011].

Neste trabalho, usamos um conjunto de dados com informações sobre temperatura do ar, como um caso de teste, mas com o *software* é possível criar outras tabelas com diferentes conjuntos de dados disponíveis no site da NOAA, como, precipitação, umidade, pressão atmosférica, entre outros. A consulta a seguir recupera as estatísticas da temperatura do ar para a região do "Estado de São Paulo" em um período do ano (Jan-Dez, 1948). Os resultados estatísticos da consulta são apresentados na Tabela 1 fornecendo a temperatura média do ar, mínimo, máximo, desvio padrão e a variância.

```
SELECT avg(air) temperatura_media, min(air) temperatura_min, max(air) temperatura_max,
stddev(air) desvio_padrao, variance(air) variância

from noaa_point, noaa_data, uf

where (uf.nome = 'SP')
      AND st_within(st_setsrid(noaa_point.lon_lat,4326),uf.the_geom)
      AND noaa_point.id = noaa_data.point_id
      AND noaa_data.time_line between '1948/01/01' and '1948/12/01'
```

Tabela 1. Estatísticas da temperatura do ar para "São Paulo" (Jan-Dec, 1948).

temperatura_media	temperatura_min	temperatura_max	desvio_padrao	variância
20.66	15.54	29.32	2.14	4.58

Este resultado é apenas um exemplo do que poderia ser feito a partir dos dados armazenados em um GIS. Diversas outras consultas e análises na base de dados foram realizadas, entretanto não serão qui apresentadas pela restrição de espaço.

Para futuros trabalhos, outras consultas SQL relativamente simples podem se realizadas, como por exemplo, *"Dada uma geometria representada por um rio ou uma estrada, qual é a temperatura média em um dado período de tempo?"* Além disso, podem ser exploradas estatísticas avançadas como a correlação e a autocorrelação, como por exemplo, *avaliar a semelhança dos padrões da temperatura do ar na região do "Vale do Paraíba Paulista" para dois períodos de anos diferentes*. O trabalho futuro empregará técnicas de DM e Reconhecimento de Padrões para descobrir conhecimento fundamental na massa de dados para executar operações mais complexas em classe similares de tabelas e geometrias.

5. Conclusões

A análise de dados meteorológicos e climáticos normalmente começa com muitos procedimentos manuais tediosos por causa do pré-processamento dos dados para posterior análise. O *software* desenvolvido automatiza algumas atividades de pré-processamento do arquivo de dados netCDF fornecidos pelo site da NOAA e exporta esses dados para um SGBD. A potencialidade do resultado obtido no estudo de caso foi satisfatória, pois mostra que os dados armazenados em PostGIS podem ser tratadas com simples consultas SQL. Acreditamos que este trabalho será de grande valia para os outros pesquisadores que trabalham com dados no formato netCDF, pois há um aumento na produtividade em procedimentos antes manualmente executados. Além disso, permite aplicar diferentes ferramentas sobre os dados armazenados no PostGIS adaptando-os conforme a técnica de DM a ser utilizada.

6. Referências

ALMEIDA, E. S.; VINHAS, L.; CORREA, P. C. P. **Extração e manipulação do dados de reanalise do ECMWF utilizando METVIEW.** In: CONGRESSO BRASILEIRO DE METEOROLOGIA, 10., 1998, Brasilia. **Anais...** 1998. CD-ROM. (INPE-10917-PRE/6373). Disponível em: <http://urlib.net/cptec.inpe.br/walmeida/2004/07.07.13.02>. Acesso em: 30 set. 2011.

CÂMARA, G. CASANOVA, M. A. HEMERLY, A. S. MAGALHÃES, G. C. MEDEIROS, C. M. B. **Anatomia de Sistemas de Informação Geográfica.** X Escola de Computação, Instituto de Computação, UNICAMP, 1996.

ESRL. ***Earth System Research Laboratory. Physical Sciences Division.*** Disponível em: <http://www.esrl.noaa.gov>. Acesso em: 15 jul. 2011.

IBGE. **Instituto Brasileiro de Geografia e Estatística.** Disponível em: <http://www.ibge.gov.br >. Acesso em: 15 agost. 2011.

NOAA. ***About NOAA.*** *National Oceanic and Atmospheric Administration. United States Department of Commerce*. Disponivel em: <http://www.noaa.gov/about-noaa.html >. Acesso em: 15 jul. 2011.

UNIDATA. ***Unidata Overview.*** http://www.unidata.ucar.edu/publications/directorspage/UnidataOverview.html. Acessado em 14 de jul de 2011.