Mineração de Dados Complexos

INF-0611 - Dados Complexos e Recuperação de Informação

Trabalho Final da Disciplina

Trabalho em Dupla

Contexto e Base de Dados

Tal como você aprenderá ao longo das demais disciplinas do curso, um dos desafios da área de Mineração de Dados Complexos é transformar um conjunto de dados denso e pouco compreensível (à primeira vista) em uma história que faça sentido em um contexto específico. Assim como em narrativas, uma sequência cronológica de fatos é importante para o entendimento do enredo, assim como é importante relatar como os dados foram coletados, manuseados e representados até que as descobertas (informação) possam se tornar viáveis.

A base de dados climáticos do CEPAGRI (Centro de Pesquisas Meteorológicas e Climáticas Aplicadas à Agricultura) da Unicamp será seu ponto de partida para sua aventura no mundo dos dados e das séries temporais em que aplicará seu conhecimento obtido até aqui. Esta base possui dados das seguintes variáveis temporais e climáticas: **data e hora**, **temperatura**, **velocidade do vento**, **umidade relativa** e **sensação térmica**. Trata-se, portanto, de uma série temporal unidimensional e multivariada. Para obtê-la, acesse o seguinte endereço Web:

http://www.ic.unicamp.br/~zanoni/cepagri/cepagri.csv

A tarefa (história) que propomos a você está inserida no contexto de Recuperação da Informação. A ideia é que você aplique o que foi visto em sala de aula para criar sua solução que possibilite a **Busca e Recuperação de Informação** relevante da base. Nesse sentido, a primeira etapa consiste no **Pré-Processamento** do conjunto de dados disponibilizado, na qual você limpará e estruturará os dados. Essa etapa também será contemplada no trabalho final da disciplina INF-0612 (Análise de Dados), então além de conhecer os seus dados você usará a mesma base pré-processada para o trabalho desta disciplina. Você notará, por exemplo, que alguns dados encontram-se ausentes para determinados dias. Caberá a você lidar com tais circunstâncias da maneira que achar mais coerente, porém, sempre respeitando as boas práticas de tratamento de dados. Assim como na outra disciplina, use os dados do período *de 01/01/2015 a 31/12/2017*.

Enunciado

Objetivo

Projetar e implementar uma solução de busca de informação em que, dada como entrada uma série de medições referentes a um dia (chamaremos de *query*), recuperar os dias, cujas séries de medições são relevantes para aquela *query*.

Sua solução deverá recuperar as 100 (cem) séries mais relevantes para a consulta estipulada. O escopo deste trabalho é na camada de processamento de um sistema de recuperação de informação.

Roteiro

No Diagrama 1 ilustramos o fluxo de trabalho esperado para a resolução deste trabalho.

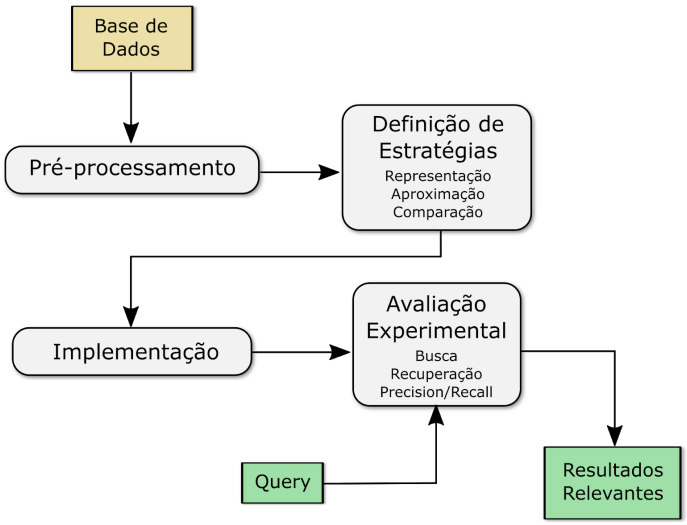

Diagrama 1: Fluxo de trabalho.

Após um **Pré-Processamento** inicial dos dados, temos a etapa de **Definição de Estratégias** de representação, aproximação e comparação dos dados, a fim de pavimentar a busca e a sua futura recuperação. Nesta etapa, você deverá tomar algumas decisões, tais como:

- Como representarei os meus dados?
- Qual será o meu descritor? Usarei um descritor específico para extrair informações de interesse (vetor de características)? Que atributos representam melhor meus dados?
- Adotarei quais métricas de similaridade/distância para comparar duas séries de dados?
- Como modelar uma série temporal? Como modelar a série de dados de outra forma?
- Com qual resolução dos dados trabalharei?

- Preciso fazer alguma transformação em meus dados? Precisarei usar o SAX para reduzir a dimensionalidade da série e obter uma representação simbólica? Dynamic Time Warping (DTW) funcionaria melhor? É necessário normalização?

Requisito: você deve projetar **duas estratégias**, uma explicitamente com uso de **séries temporais e uma outra projetada de forma diferente**. Para cada estratégia considere duas medidas de distância/similaridade diferentes.

Definida as estratégias, chegamos à etapa de **Implementação**. Neste momento, você deverá implementar, em Linguagem R, a solução para o objetivo definido neste trabalho.

Importante: os dados de data servirão apenas para avaliação dos métodos propostos e ordenação de dados da séries durante o desenvolvimento do seu método. Portanto, **não baseie a recuperação de séries no atributo data**. Apesar de o tempo ser um fator relevante no âmbito de séries temporais, neste trabalho estamos em um cenário de busca em que a *query* não se valerá da data (apenas usada na avaliação do sistema).

Finalmente, para a etapa de **Avaliação** da sua solução, devemos definir o que é relevante e o que não é, a fim de avaliar o resultado da sua busca/recuperação por meio das métricas precisão (precision) e revocação (recall).

Neste trabalho, usaremos o atributo data para definir o que é relevante ou não (***ground truth***). Considere que as séries relevantes para uma dada *query* são aquelas cujas datas estão dentro do intervalo dos **7 dias anteriores e 7 subsequentes à data da consulta (*query*)**.

A fim de avaliar a eficiência do seu método, você deve realizar experimentos com 100 *queries* aleatoriamente selecionadas da base de dados. **Baixe** do Moodle e **use** o arquivo **query.csv,** que contém as datas das séries que devem ser usadas como *query* de avaliação dos seus métodos. Ou seja, dada uma data X, a query será composta por todas as medições referentes à data X.

Você deve relatar a **revocação (recall) e precisão (precision) médias** das 100 *queries* usando P@K, sendo K = {5, 10, 15, 20, 25, 30, …, 100}, ou seja, analisará a precisão média dos K primeiros itens da lista retornada pela sua solução/método e sua respectiva revocação média.

Considere mostrar:

- Tabelas com a **média** dos valores de precisão e revocação com P@K das 100 *queries*;
- Gráfico comparando as 4 curvas de precisão-revocação (dois métodos propostos, cada um usando duas funções de distância/similaridade distintas);
- Outras análises que julgar relevante.

Protocolo de Submissão dos Resultados

Você deve submeter um arquivo compactado (formato zip), contendo os seguintes documentos/arquivos:

- Código R referente às implementações realizadas, devidamente comentado;
- Relatório técnico (em PDF) contendo, nome completo dos integrantes da dupla e breve descrição, dentre outros assuntos, sobre eventuais procedimentos de pré-processamento realizados, suas decisões de projeto, os algoritmos efetivamente implementados, eventuais dificuldades enfrentadas, os resultados e as suas análises.