UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

CONSULTAS POR SIMILARIDADE E MINERAÇÃO DE REGRAS DE ASSOCIAÇÃO: MAXIMIZANDO O CONHECIMENTO EXTRAÍDO DE SÉRIES TEMPORAIS

CLAUDINEI GARCIA DE ANDRADE

ORIENTADORA: PROFA. DRA. MARCELA XAVIER RIBEIRO

São Carlos – SP

Julho/2014

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

CONSULTAS POR SIMILARIDADE E MINERAÇÃO DE REGRAS DE ASSOCIAÇÃO: MAXIMIZANDO O CONHECIMENTO EXTRAÍDO DE SÉRIES TEMPORAIS

CLAUDINEI GARCIA DE ANDRADE

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação, área de concentração: Engenharia de Software / Banco de Dados
Orientadora: Profa. Dra. Marcela Xavier Ribeiro

São Carlos – SP

Julho/2014

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

A553cs

Andrade, Claudinei Garcia de.
Consultas por similaridade e mineração de regras de associação : maximizando o conhecimento extraído de séries temporais / Claudinei Garcia de Andrade. -- São Carlos : UFSCar, 2014.
67 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2014.

1. Data mining (Mineração de dados). 2. Análise de séries temporais. 3. Regras de associação. 4. Consultas por similaridade. 5. Coulomb, Lei de. I. Título.

CDD: 005.741 (20ª)

Universidade Federal de São Carlos

Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Ciência da Computação

"Consultas por Similaridade e Mineração de Regras de Associação: Maximizando o conhecimento extraído de Séries Temporais "

Claudinei Garcia de Andrade

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação

Membros da Banca:

Profa. Dra. Marcela Xavier Ribeiro
(Orientadora - DC/UFSCar)

Profa. Dra. Luciana Alvim Santos Romani
(EMBRAPA Informática Agropecuária – Campinas)

Prof. Dr. José Fernando Rodrigues Júnior
(ICMC/USP)

São Carlos
Agosto/2014

Dedico este trabalho à minha filha
Sophia.

AGRADECIMENTOS

Primeiramente a Deus, pelo dom da vida e pela graça de todos os dias encontrar motivos para amá-Lo mais em todas as coisas e, a cada descoberta, poder amá-Lo mais que todas as coisas por meio das quais Ele se revela.

Aos meus pais, por terem me dado a riqueza de conhecimentos mais valiosos do que o melhor dos resultados que esta dissertação puder conseguir. Pelo carinho incondicional e pela perseverança em amar sempre.

Ao Programa de Pós-Graduação em Ciência da Computação da UFSCar, por possibilitar um grande enriquecimento intelectual permitindo o desenvolvimento deste trabalho. Aos docentes do programa, pela dedicação ao ensino de qualidade.

À professora Dra. Marcela Xavier Ribeiro, pela amizade e confiança adquiridas durante esses anos de convivência e por todo apoio para a realização deste trabalho.

Aos funcionários e alunos do DC-UFSCar, pela saudável convivência e por todo apoio dado durante a realização deste trabalho.

A todos que, direta ou indiretamente, colaboraram na execução deste trabalho, em especial à Mirela Cazzolato, profa. Elaine Parros, Luciana Alvim e Renata Ribeiro pelas sugestões e pelo apoio.

À Justiça Federal de Primeiro Grau em São Paulo pelo apoio e incentivo.

Verdadeiramente o que mais prazer me proporciona,
não é o saber, mas o estudar,
não a posse, mas a conquista,
não o estar aqui, mas o chegar além.

Carl Friedrich Gauss

RESUMO

A análise de séries temporais apresenta certos desafios. Seja pela dificuldade na manipulação dos dados, por exigir um grande custo computacional, ou mesmo pela dificuldade de se encontrar subsequências que apresentam as mesmas características. No entanto, essa análise é importante para o entendimento da evolução de diversos fenômenos como as mudanças climáticas, as variações no mercado financeiro entre outros. Este projeto de mestrado propôs o desenvolvimento de um método para a realização de consultas por similaridade em séries temporais que apresentam melhor desempenho e acurácia que o estado-da-arte e um método de mineração de regras de associação em séries utilizando similaridade. Os experimentos feitos aplicaram os métodos propostos em conjuntos de dados reais, trazendo conhecimento relevante, indicando que os métodos são adequados para análise por similaridade de séries temporais unidimensionais e multidimensionais.

Palavras-chave: mineração de dados, séries temporais, regras de associação e consulta por similaridade.

ABSTRACT

A time series analysis presents challenges. There is a difficulty to manipulate the data by requiring a large computational cost, or even, by the difficulty of finding subsequences that have the same characteristics. However, this analysis is important for understanding the evolution of various phenomena such as climate change, changes in financial markets among others. This project proposed the development of a method for performing similarity queries in time series that have better performance and accuracy than the state-of-art and a method of mining association rules in series using similarity. The experiments performed have applied the proposed methods in real data sets, bringing relevant knowledge, indicating that both methods are suitable for analysis by similarity of one-dimensional and multidimensional time series.

Keywords: data mining, time series, association rules, similarity search

LISTA DE FIGURAS

LISTA DE TABELAS

GLOSSÁRIO

APCA – *Adaptive Piecewise Constant Approximation*

DFT – *Discrete Fourier Transform*

DWT – *Discrete Wavelet Transform*

FFT – *Fast Fourier Transform*

FM – *Flexible Module*

PAA – *Piecewise Aggregate Approximation*

PCA – *Principal Component Analysis*

PLA – *Piecewise Linear Aproximation*

SAX – *Symbolic Aggregate Approximation*

SM – *Sequential Matching*

SVD – *Singular Value Decompositon*

TSS – *Tractable Similarity Searching*

k-NN query – *k-Nearest neighbor query*

SUMÁRIO

Capítulo 1

INTRODUÇÃO

Este capítulo apresenta o contexto, motivação e os objetivos deste trabalho de mestrado, sendo dividido da seguinte maneira: Seção 1.1 apresenta as considerações iniciais sobre o projeto e o contexto no qual o presente trabalho se insere. Seção 1.2 apresenta a motivação para a realização deste projeto, indicando a contribuição do trabalho em questão. Já a Seção 1.3 descreve o objetivo principal da execução do projeto. Seção 1.4 elenca a organização desta dissertação e, finalmente, na Seção 1.5 são dadas as considerações finais deste capítulo.

1.1 Considerações Iniciais e Contexto

Desde o começo da ciência, mesmo antes da introdução dos experimentos como métodos de se replicar os fenômenos da natureza, a observação já se constituía como um dos fatores importantes para se comprovar a veracidade de um fato ou para validar alguma teoria. Atualmente, as observações são utilizadas nas mais diversas áreas do conhecimento e, juntamente com os experimentos e pesquisas de campo, permitem descobrir a relação entre elas e, assim, inferir generalizações e produzir conhecimento.

Nesse contexto, a maneira de se auferir as observações tem evoluído ao longo do tempo com uma precisão ímpar, principalmente com os sensores. Esses dispositivos são capazes de detectar mudanças nas condições de um determinado ambiente e transmitir o resultado em intervalos de tempos regulares, como uma medida ou uma instrução de controle, para uma central de gerenciamento. Esses conjuntos de observações tomados no decorrer de intervalos de tempo são conhecidos como séries temporais.

Com o avanço tecnológico, associado ao baixo custo da produção de instrumentos para mensurar observações, tem crescido vertiginosamente a quantidade de dados disponíveis para

análise. No entanto, os dados coletados, na sua grande maioria, apresentam relações intrínsecas entre eles que não são perceptíveis sem uma análise minuciosa. Necessitando, assim, da utilização de técnicas específicas para se conseguir obter conhecimento a partir destes dados.

Assim, o desenvolvimento de técnicas computacionais efetivas e eficazes para a análise de séries temporais se faz necessário.

1.2 Motivação

Dados de séries temporais são geralmente provenientes de sensores, o que gera uma grande quantidade de dados para análise. E a análise desses dados, torna a manipulação de séries temporais muito custosa, pois demanda a análise temporal simultânea de diversas observações e requer comparações exaustivas dos elementos desse conjunto. Assim, um dos grandes desafios iniciais é o desenvolvimento de métodos que permitam a redução da dimensionalidade da série, de maneira que a manipulação delas seja menos custosa e ao mesmo tempo mantendo-se a precisão com relação às características dos dados e sem muita perda de informação.

Outro fator importante a se levar em consideração é que as séries temporais são consideradas dados complexos. E nesses tipos de dados, a consulta mais adequada é a consulta por similaridade, pois não existe relação de ordem total entre os elementos do conjunto e a obtenção de elementos iguais é muito difícil de acontecer. No entanto, a obtenção da similaridade entre séries ou subsequências de uma série também não ocorre de forma trivial. A busca por similaridade exige, na maioria dos casos, a utilização de descritores que pode exigir alto custo computacional para o processamento das consultas e que podem não apresentar resultados satisfatórios.

Além disso, uma técnica muito utilizada para mineração de séries temporais é a discretrização dos dados, realizando comparações de casamento exato, como dados convencionais. Esse tipo de abordagem tende a reduzir o potencial de padrões que podem ser minerados a partir das séries. Assim, é importante o desenvolvimento de métodos de mineração com base em comparações por similaridade entre os dados para uma melhor identificação de padrões. Logo, o presente trabalho foi motivado por desenvolver métodos que integrem mineração de regras de associação e consultas por similaridade para a análise e obtenção de conhecimento em séries temporais.

1.3 Objetivos

1.4 Organização do Trabalho

O objetivo deste trabalho foi desenvolver métodos de redução de dimensionalidade das séries através do desenvolvimento de um descritor que permite a realização de consultas por similaridade em séries temporais uni e multidimensionais e que apresente um desempenho e acurácia melhores que o estado da arte. E também, o desenvolvimento de um método para a mineração de regras de associação que emprega o descritor proposto para encontrar fenômenos não evidentes em dados meteorológicos.

1.5 Considerações Finais

Neste capítulo foi apresentada uma introdução ao projeto de pesquisa, mostrando a importância da análise das séries temporais e as dificuldades encontradas para a execução deste trabalho. Assim, a utilização de descritores torna mais robusto o processo de busca de similaridade em séries temporais e também a mineração dos dados, pois isso facilita a indexação para pesquisas e, por consequência, a extração de conhecimento intrínseco, deixando claras a necessidade e importância da execução do presente trabalho.

Capítulo 2

TÉCNICAS DE EXPLORAÇÃO DE SÉRIES TEMPORAIS

Para a análise de séries temporais, um dos desafios é encontrar uma maneira de representá-las de maneira precisa para que as comparações entre séries ou subsequências da série possam ser executadas de maneira ágil. Neste capítulo serão discutidas maneiras de analisar séries. Na Seção 2.1 são introduzidas as considerações iniciais. A Seção 2.2 descreve a análise de séries temporais. A Seção 2.3 explana sobre a consulta por similaridade em séries e na Seção 2.4 são elencados os principais descritores existentes na literatura para séries. Além disso, a Seção 2.5 define os conceitos acerca de funções de distância, sendo estas necessárias para a verificação da distância e da similaridade entre séries. Na Seção 2.6, as técnicas de validação para análise de séries temporais são mostradas e na Seção, 2.7, é discutida a mineração de regras de associação. E por fim, na Seção 2.8, são feitas as considerações finais deste capítulo.

2.1 Considerações Iniciais

Para a extração de conhecimento contido em séries temporais, a análise da série exige a observância de certas características intrínsecas a elas, pois a utilização de um dado ou de uma determinada subsequência de maneira isolada, em geral, não é suficiente para representar a série como um todo. E ainda, a representação de uma série ou de uma subsequência da série por meio de uma maneira compacta, também pode causar distorções no processo de mineração.

Assim, a representação adequada da série de maneira que facilite a extração de conhecimento e que torne fácil a sua manipulação computacional e, ainda, que preserve o máximo das informações originais constituem um dos pilares para a análise de séries (BARIONI, 2006).

Neste capítulo, será abordado o ferramental para consultas por similaridade relacionado ao processo de extração de conhecimento em séries temporais, bem como, apresentadas as principais técnicas existentes para execução desse processo.

2.2 Análise de séries temporais

Pela definição clássica de série temporal a ordenação em função do tempo das observações é muito importante, no entanto, não é somente o tempo que pode ser considerado um índice para as aferições. Numa descrição mais genérica, uma série temporal pode ser definida como uma sequência ordenada de observações (WEI, 2006). O índice utilizado para ordenar essa sequência pode ser o tempo ou outro qualquer como: espaço, profundidade, entre outros.

Formalmente, uma série temporal unidimensional é um conjunto de observações $\{Y(t), t \epsilon T\}$ em que Y é a variável de interesse e T é conjunto de índices. Uma subsequência de uma série Y de tamanho m pode ser definida como $\{Y(t), t \epsilon T\}$ em que $\{1 \leq t \leq m\}$. Uma série multidimensional Y_m de tamanho n é uma sequência de m conjuntos de valores (TANAKA; IWAMOTO; UEHARA, 2005), representada por $Y_m = (x_{11}, ..., x_{m1}), ..., (x_{1n}, ..., x_{mn})$.

Podemos classificar as séries em 3 tipos básicos com relação ao intervalo de observações (WEI, 2006). Sendo: i) série discreta, se as observações são feitas em tempos determinados e, geralmente, regulares $T = \{t_1, t_2, ..., t_n\}$; ii) série contínua, quando as observações são contínuas no tempo e $T = \{t : \ t_1 < t < t_2\}$; e iii) multivariadas, se apresentam várias observações para um mesmo tempo $Y_1(t), ..., Y_k(t), \ (t \epsilon T)$.

As séries temporais podem ser classificadas em estacionárias e não estacionárias. As estacionárias, também conhecidas como séries convergentes, permanecem em equilíbrio em torno de um nível médio constante e estão relacionadas à grande parte da teoria de séries temporais. Já as séries não estacionárias não apresentam convergência em torno de uma média. Elas também podem ser descritas utilizando seus componentes básicos. São eles: tendência, ciclo e sazonalidade (BUSSAB; MORETTIN, 2008).

A análise da tendência em uma série consegue indicar o seu comportamento em um período relativamente longo de tempo. Isso ocorre devido ao fato de que é necessária uma grande quantidade de dados para representar a série e realizar os cálculos e, a partir daí, verificar se ela cresce, decresce ou permanece estável, e também qual a velocidade dessas mudanças. Pode-se, também obter a função geradora da tendência e assim realizar um estudo mais detalhado.

Outro componente importante para análise de séries é a presença de ciclos em suas com-

ponentes. O ciclo pode ser caracterizado pelo movimento oscilatório de grande duração ao longo da série, fazendo com que a série apresente uma variação que se repete, mas que não está associada automaticamente a nenhuma medida temporal.

A sazonalidade, outro componente importante das séries, está ligada às variações periódicas, da mesma forma que os ciclos, no entanto, ocorrem em intervalos regulares.

Assim, com a análise dos componentes e características da série é possível fazer uma análise do conteúdo da mesma tendo como objetivos:

- Descrever a série mostrando as propriedades constitutivas dela como tendência, sazonalidade entre outras;
- Compreender o mecanismo da série possibilitando encontrar razões para o comportamento dela;
- Predizer valores futuros, utilizando dados e comportamentos passados e também métodos de previsão; e
- Obter controle sobre o processo que gera as observações e, assim, garantir que a série tenha um comportamento já esperado.

Obtendo as características pertinentes à série, pode-se descobrir e visualizar padrões nas séries, detectar anomalias, identificar séries ou intervalos semelhantes, gerar agrupamentos, regras de associação, entre outras atividades em que as características obtidas da série possam ser utilizadas como norteadoras de identificação.

Um fator importante a se considerar na análise de séries é a redução da dimensionalidade. Uma série temporal pode ser considerada uma sequência de dados, em que, a cada ponto é atribuído uma dimensão (ou comprimento) n e que reduzi-la para uma dimensão k, com $k \prec\prec n$, implica em reduzir o custo computacional para consultas em séries temporais.

2.3 Consulta por similaridade em séries

Séries temporais são consideradas dados complexos, que são dados que não apresentam relação de ordem total e, logo, não existem maneiras triviais de se estabelecer uma relação de ordem entre séries ou suas subsequências. Além disso, devido à grande variabilidade existente nos dados é quase impossível encontrar séries ou intervalos iguais. Nesse contexto, o conceito de

similaridade tem maior aplicabilidade que o conceito de igualdade, pois a consulta por similaridade, feita especificamente para este domínio, retorna objetos do conjunto de dados que sejam similares a um objeto de consulta, ocasionando melhores resultados que a busca por igualdade (BARIONI, 2006)

Para a execução das consultas por similaridade é necessário haver um meio de mensurar a similaridade ou de dissimilaridade existente entre dois objetos pertencentes ao domínio.

Um espaço métrico M pode ser definido pelo par $\{\mathcal{S}, d\}$, em que $\mathcal{S}$ define o domínio dos dados e d é uma função de distância que fornece uma medida de quão similar ou dissimilar um objeto é do outro (BOZKAYA; OZSOYOGLU, 1999)

No entanto, para a aplicação de funções de distância em dados complexos, nem sempre é possível ou viável utilizar os dados propriamente ditos. Uma alternativa comumente utilizada é a extração de características inerentes a esses dados, sendo que cada característica é um valor ou um conjunto de valores numéricos e o conjunto dessas características extraídas formam um vetor de características.

O vetor de características é utilizado pelas funções de distância para o cálculo da similaridade e, consequentemente, para as operações de busca e comparação dos dados, retornando como resultado da consulta um conjunto de objetos similares ordenados pela similaridade em relação ao objeto de referência. Essa abordagem é chamada de recuperação por conteúdo.

Existem dois tipos básicos de consultas por similaridade: i) a consulta por abrangência e ii) a consulta aos k-Vizinhos mais próximos:

- Consulta por abrangência (*Range query*): visa encontrar todos os objetos pertencentes ao domínio que sejam dissimilares ou similares de um objeto de consulta Q até no máximo certo limitante r. Ou utilizando outra abordagem, a consulta por abrangência visa encontrar objetos que estejam a uma distância no máximo r do objeto de consulta Q, conforme ilustra a Figura 2.1. No caso de séries temporais, dado uma subsequência pertencente a uma série, uma consulta por abrangência deve retornar as subsequências com maior similaridade dentro de uma distância máxima r do objeto de consulta;

- Consulta aos k-Vizinhos mais Próximos (*k-Nearest Neighbor query* ou *k-NN query*): visa recuperar os k objetos mais semelhantes a um objeto de consulta, conforme ilustrado na Figura 2.2. No caso de séries temporais, uma consulta aos k objetos mais similares a uma subsequência deve retornar as k subsequências mais similares pertencente à série temporal.

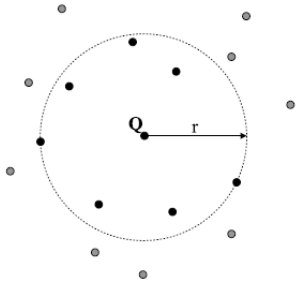

Figura 2.1: Exemplo de uma consulta por abrangência a um objeto Q e utilizando uma distância máxima r.

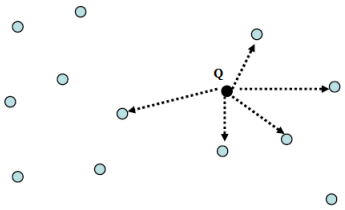

Figura 2.2: Exemplo de uma consulta k-NN com $k = 4$ a um objeto de referência Q.

Para séries temporais, podem ser executadas consultas de duas maneiras distintas (SANTOS, 2011):

- **Busca por subsequências**: dada uma série temporal e uma subsequência pertencente à própria série, a consulta é realizada na série procurando por subsequências similares ao intervalo dado; e
- **Busca por uma série inteira**: neste caso, uma série é passada como consulta e outras séries similares pertencentes ao conjunto de busca são retornadas, caso haja similaridade entre elas.

2.4 Descritores de séries

Na literatura, não há uma consolidação sobre o conceito de descritor para dados complexos. Alguns autores (TORRES; FALCÃO, 2006) definem um descritor como sendo formado por uma tupla (ϵ_D, δ_D) em que:

- $\boldsymbol{\epsilon_D}$: é o componente responsável por caracterizar o objeto, por meio da extração de características e gerando um vetor que servirá para analisar os dados; e
- $\boldsymbol{\delta_D}$: é a função responsável por comparar os vetores de características, dando a quantidade de similaridade existente entre o objeto e a consulta.

No entanto, encontra-se na literatura, o conceito de descritor se referindo somente à função que gera o vetor de características. Neste trabalho, será referenciado como descritor o conjunto formando pelo vetor de característica e a função de distância.

A seguir é apresentada a descrição dos principais métodos utilizados na busca por similaridade em séries.

2.4.1 *Sequential Scan*

O método *Sequential Scan*, também conhecido como solução de Força Bruta, *Sequential Matching* ou *Sequential Scanning* é citado em: (FALOUTSOS; RANGANATHAN; MANOLOPOULOS, 1994) e (KEOGH, 1997), e é considerado um método trivial para busca de similaridade em série. Ele consiste basicamente em deslocar uma subsequência de consulta ao longo de toda a série calculando a distância. Geralmente é utilizada a função de distância quadrática, entre cada observação e buscando sequencialmente toda possível subsequência pertencente a sequência que seja o mais similar possível com a consulta inserida.

Formalmente, dada uma série Y de tamanho m e uma consulta Q de tamanho n, o descritor por força bruta busca a solução que minimize a Equação 2.1:

$$\min_{1 \leq i \leq m} \sum_{t=1}^{n} (Y_i(t) - Q(t))^2 \tag{2.1}$$

Este método apresenta como vantagens o fato de ser um ótimo método para busca por similaridade. No entanto, sua desvantagem é a alta complexidade computacional. A complexidade deste método é $\mathbf{O}(m-n+1) * n$ (KEOGH, 1997) onde m é o número de pontos da série pesquisada e n é o número de pontos existentes na consulta. Logo, para uma série que apresenta grande quantidade de pontos fica inviável a sua aplicação.

2.4.2 Transformada Discreta de *Fourier*

O descritor baseado na Transformada discreta de *Fourier* ou *Discret Fourier Transform* - DFT é um método baseado em processamento de sinais de *Joseph Fourier* em que uma equação pode ser expressa como uma combinação linear de soluções harmônicas. O descritor que utiliza DFT foi proposto por Agrawal, Faloutsos e Swami (1993) em que, segundo os autores, um pequeno número de coeficientes é suficiente para uma boa descrição para a grande maioria das funções. E este foi um dos primeiros métodos propostos para a redução de dimensionalidade em séries e para a busca de similaridade em séries.

A transformada de *Fourier* apresenta uma grande quantidade de variantes e para o estudo de séries é utilizada a transformada rápida de *Fourier* que apresenta um custo computacional menor, $\mathbf{O}(nlog(n))$, se comparado com a técnica original, $\mathbf{O}(n^2)$, com n representando o tamanho da entrada. Além disso, a função de distância comumente utilizada é a *L2*.

Dado uma subsequência da série de tamanho n formado por $(x_1, x_2, ..., x_n)$, a transformada

rápida de *Fourier* (*Fast Fourier Transform* - FFT) reduz a dimensionalidade da subsequência representando-a por X_k utilizando a Equação 2.2:

$$X_k = \sum_{n=0}^{N-1} (x_n.e)^{-i2\pi \frac{k}{N} n} \tag{2.2}$$

Por ser uma transformação que expressa uma série temporal em termos de uma combinação linear de base sinusoidal, ela é muito eficiente para determinar o espectro de frequência de um sinal, ou seja, para a determinação de pontos de inflexão na série. No entanto, para a análise de séries temporais estacionárias, em que a variação dos valores é pequena, o resultado obtido pela representação da série por FFT também apresenta uma pequena variação e isso pode dificultar a análise da série.

2.4.3 Decomposição em valores singulares

Esse descritor conhecido como *Singular Value Decomposition* - SVD, proposto por Korn, Jagadish e Faloutsos (1997) é a representação da série por uma combinação linear de formatos, ou seja, a série é representada por uma matriz **A** de tamanho *m* x *n* e o descritor SVD de **A** é dado pela equação 2.3:

$$A_{mxn} = U_{mxn} S_{nxn} V^T_{nxn} \tag{2.3}$$

Em que **S** representa um vetor com os autovalores de **A**. As matrizes **U** e **V** são as decomposições de uma base ortonormal para as colunas e linhas de **A**, respectivamente.

Este método apresenta como vantagem representar a série sem grandes perdas de dados se comparado ao descritor DFT, no entanto, o cálculo de autovetores e autovalores tem um grande custo computacional. Para a representação de subsequências grandes, a redução de dimensionalidade apresenta perdas. A função de distância comumente utilizada para esse vetor de características é a *L2*.

2.4.4 Transformada discreta de *wavelet*

O *Discrete Wavelet Transform* - DWT, proposto por Chan e Fu (1999) transforma a série em uma combinação linear de funções com base na definição de *wavelet* do matemático A. Haar.

A DWT, utilizada em séries, baseia-se em uma adaptação do conceito de *wavelet* de Haar,

onde há duas funções: uma **função de translação** que transforma os dados da série para a aplicação da **função de escala**. Essa, por sua vez, transporta os dados para um intervalo que varia de -1 a 1 e reduz a dimensionalidade dos dados.

Esse descritor apresenta-se ineficiente para a representação de dados que apresentem grandes amplitudes ou uma grande variabilidade dos dados, pois há uma supressão de características importantes no momento da transformação da função de translação para a função de escala. Ou seja, a DWT consegue captar variações nos dados de acordo com o que foi definido na função de translação e não para qualquer intervalo da série.

De maneira geral, os três descritores anteriormente apresentados utilizam técnicas baseadas em processamento de sinal e são bastante utilizados na busca de similaridade em séries. Além disso, esses descritores transformam o conjunto de dados em um pequeno subconjunto de coeficientes que são considerados representantes para determinado intervalo da série. Eles tendem a ser eficientes para o processamento computacional, entretanto, para longas séries temporais, a redução de dimensionalidade pode representar de maneira insatisfatória a série.

Outros descritores foram propostos na literatura, não utilizando a abordagem de processamento de sinais, sendo eles:

- ***Piecewise aggregate approximation*** - PAA: proposto por Keogh et al. (2001) e representa a série por meio de uma sequência de segmentos de igual tamanho, utilizando para isso o valor médio da subsequência e a distância utilizada, geralmente, é *L1*;
- ***Adaptive Piecewise Constant Approximation*** - APCA (CHAKRABARTI et al., 2002): esse descritor é um melhoramento do descritor PAA, em que os segmentos apresentam tamanhos adaptativos e são apresentados vários segmentos em períodos da série que apresentam grande variabilidade e poucos segmentos em intervalos de baixa variabilidade. A distância utilizada, geralmente, é *L1*;
- ***Piecewise Linear Aproximation*** - PLA (MORINAKA et al., 2001): esse descritor representa a série por uma sequência de linhas retas e o cálculo da função de distância se baseia no comprimento da linha e da altura em que ela se encontra;
- ***Symbolic Aggregate Approximation*** - SAX proposto por Lin et al. (2003) e melhorado por (CAMERRA et al., 2010). Esse descritor converte a série em uma sequência de caracteres de acordo com a variabilidade dos dados e utiliza uma função de distância baseada em texto para o cálculo de similaridade.

- ***Dynamic time warping*** - DTW proposto por (BERNDT; CLIFFORD, 1994) é um descritor que usa uma função de distância baseada em segmentos não lineares entre séries temporais ou subsequências de séries para o cálculo da dissimilaridade. Ele é considerado muito eficaz para classificação de séries por ter uma boa capacidade de adaptação para tratar de desvios que ocorrem no eixo do tempo. No entanto, o cálculo da distância entre os segmentos tem um grande custo computacional (custo quadrático).

2.5 Funções de distância para séries

O vetor de características tem importância fundamental para a busca por similaridade em séries, no entanto, ele não é completamente suficiente para a análise de similaridade em séries. Para isso se faz necessário comparar esses vetores por meio de uma função que avalie o quão similar ou dissimilar um vetor de características é de outro. Dá-se o nome de função de distância ou função de similaridade a essa função.

A função de distância deve respeitar as propriedades inerentes ao espaço métrico M. Dados dois elementos, x e y, pertencentes a um domínio e sendo d uma função de distância entre eles, as propriedades devem ser válidas:

- **Simetria**: $\forall x,\ y \in M,\ d(x,y) = d(y,x)$;
- **Não-negatividade**: $\forall x,\ y \in M,\ x \neq y, d(x,y) > 0$ e $d(x,x) = 0$;
- **Desigualdade triangular**: $\forall x,\ y \in M,\ d(x,y) \leq d(x,z) + d(z,y)$.

Há várias funções de distância, principalmente na área de imagens e cada uma delas pode ter um melhor resultado se aplicada a um domínio específico. As principais funções de distância são conhecidas como funções de distância de *Minkowski* (família L_p). As principais são:

2.5.1 Distância de *Manhattan*

A função de distância de *Manhattan* ou $L1$, também chamada de distância de *city-block*, é uma função simples e bastante utilizada. Seu funcionamento consiste em: dados dois vetores de características $X = \{x_1, x_2, ..., x_n\}$ e $Y = \{y_1, y_2, ..., y_n\}$, a distância é calculada pela soma das diferenças entre os módulos dos elementos correspondentes, conforme mostra a Equação 2.4:

$$L_1(X,Y) = \sum_{k=1}^{n} (|x_k - y_k|) \tag{2.4}$$

Onde n é o tamanho do vetor de características.

2.5.2 Distância Euclidiana

A função de distância Euclidiana ou $L2$, também conhecida como distância quadrática, consiste em calcular a distância entre dois vetores de características $X = \{x_1, x_2, ..., x_n\}$ e $Y = \{y_1, y_2, ..., y_n\}$, usando a diferença quadrática entre os módulos dos elementos correspondentes, conforme mostra a Equação 2.5:

$$L_2(X,Y) = \sqrt{\sum_{k=1}^{n}(|x_k - y_k|)^2} \tag{2.5}$$

Onde n é o tamanho do vetor de características.

É importante ressaltar que existem outras funções de distância da família L_p como a $L_{infinity}$ ou mesmo outras funções não pertencentes a essa família que são aplicadas para casos específicos de vetores de características de séries. Ou, ainda, uma função de distância pode ser proposta simplesmente para atender às necessidades de cálculo de similaridade para um determinado tipo de vetor. Um exemplo desse tipo de medida pode ser encontrado em Lin et al. (2003), em que os autores propõem o método SAX para reduzir a dimensionalidade da série transformando-a em uma *string* de tamanho arbitrário. E como medida de distância é utilizada uma função que retorna a distância mínima existente entre duas palavras, conhecida como L_{edit}.

2.6 Técnicas de Validação para Análise de Séries Temporais

Dentre os descritores estudados e das técnicas de redução de dimensionalidade existentes na literatura, não há um consenso a respeito de um método de validação para a geração de métricas confiáveis que possam ser utilizadas para comparar os modelos e verificar a eficácia de cada um. Nas subseções seguintes são apresentadas as principais métricas utilizadas para avaliar descritores de séries temporais.

2.6.1 Acurácia

A acurácia é uma medida utilizada por várias áreas da ciência, destinada a mensurar a quantidade de instâncias que foram corretamente preditas a partir de uma consulta recebida de entrada (STANDARDIZATION; 69, 1994).

No caso das séries temporais, essa medida é utilizada passando-se uma subsequência de

entrada e verificando a saída dada pelo sistema para comparar se os objetos retornados representam fielmente os objetos com maior similaridade entre o objeto de consulta.

Conforme exposto anteriormente, os descritores visam reduzir a dimensionalidade das séries e, conforme as peculiaridades de cada método, eles podem representar a série toda ou parte dela de maneira imprecisa, gerando resultados insatisfatórios nas consultas. Logo, maximizar a acurácia constitui um dos desafios no trabalho de busca por similaridade em séries com grande importância para a avaliação de um descritor.

2.6.2 Complexidade computacional

Outro fator importante para validação de um método para consulta de similaridade refere-se aos requerimentos de recursos indispensáveis para que um algoritmo possa resolver um problema, ou seja, referem-se à quantidade de trabalho e/ou tempo gastos na realização de um trabalho (TOSCANI; VELOSO, 2008).

Muitos dos descritores utilizados, para efetuar a redução de dimensionalidade, efetuam cálculos de grande complexidade, como a transformada discreta de *Fourier* que utiliza números complexos em seus cálculos.

Há de se observar, também, que a função de distância consome recursos para a execução dos cálculos de distância durante a consulta por similaridade. Logo, a complexidade computacional é outro fator de grande importância para a validação de um descritor, pois um método de grande complexidade, que exige grandes recursos e leva um tempo excessivo, pode ser desconsiderado para determinados fins.

2.6.3 Precisão x Revocação

Esta técnica, muito utilizada no campo de recuperação de informação, pode ser utilizada para séries. A precisão mede a fração de objetos relevantes retornados em uma determinada consulta em relação ao total de objetos retornados. Já a revocação mede a fração de objetos relevantes retornados em uma determinada consulta em relação ao total de objetos relevantes existentes da base (PENATTI, 2009).

Formalmente, para esse projeto, temos que a revocação indica a quantidade de subsequências da série temporal relevantes provenientes da consulta (I_{RC}) em relação à quantidade total de subsequências relevantes existentes na base (I_R).

$$\text{Revocação} = \frac{I_{RC}}{I_R} \tag{2.6}$$

E a precisão indica a fração que representa o subconjunto de subsequências relevantes (I_{RC}) em relação ao conjunto total da consulta (I_C).

$$\text{Precisão} = \frac{I_{RC}}{I_C} \tag{2.7}$$

Além disso, a curva de precisão por revocação indica a variação dos valores de precisão para diferentes valores de revocação. E quanto mais alta a curva estiver, mais eficaz é um descritor. Para a utilização dessa medida em séries temporais, há somente a troca de objetos pelas subsequências da série e todo o restante continua igual, inclusive a interpretação da curva.

2.7 Mineração de Regras de Associação em Séries

O processo de descoberta de conhecimento em bases de dados *(Knowledge Discovery in Databases)* tem como objetivo a identificação de padrões em conjuntos de dados, que representem informação válida, inédita, potencialmente útil e essencialmente compreensível. Já o termo *data mining*, ou mineração de dados, refere-se a um extenso campo de pesquisa composto por um conjunto de técnicas que fazem parte de uma das etapas do processo de descoberta de conhecimento em base de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

O conhecimento é obtido pela extração de novos padrões e regras que estão implícitos em grandes quantidades de informações armazenadas nos bancos de dados de organizações, por meio da aplicação de técnicas específicas, de acordo com o tipo de conhecimento a ser minerado.

As técnicas de mineração permitem fazer uma análise antecipada dos eventos, possibilitando prever possíveis padrões, sequências, tendências e comportamentos futuros, associações, agrupamentos, classificações, hierarquias de classificação, categorizações, segmentações, facilitando, assim, o processo de decisão.

A extração de conhecimento em bases de dados é um processo composto de várias etapas, iniciando, basicamente, com a coleta de dados para o problema em pauta e finalizando com a interpretação e avaliação dos resultados obtidos.

Segundo Fayyad et al. (1996) e Fayyad, Piatetsky-shapiro e Smyth (1996), esse conjunto é composto de cinco etapas: i) seleção dos dados; ii) pré-processamento e limpeza dos dados; iii)

transformação dos dados; iv) Mineração de Dados (*Data Mining*); e v) interpretação e avaliação dos resultados, em que a etapa de mineração dos dados é subdividida em escolha das atividades ou funções, escolha de algoritmos, preparação dos dados, extração do conhecimento e pós-processamento.

Existem várias tarefas de mineração. Este trabalho foca em minerar regras de associação. O processo de mineração de regras de associação consiste em examinar os dados buscando correlações entre os seus conjuntos de itens. Ou seja, uma regra de associação mostra o quanto a presença de um conjunto de itens nos registros de uma base de dados implica na presença de algum outro conjunto distinto de itens nos mesmos registros (AGRAWAL; SRIKANT, 1994). Em suma, o objetivo do algoritmo de Regras de Associação é localizar tendências que visem entender e predizer padrões de comportamento dos dados.

A mineração de associações em uma base de dados pode gerar uma infinidade de regras sendo que algumas delas podem não ser interessantes devido à baixa frequência com que os dados aparecem, para isso são utilizados dois fatores que auxiliam no processo de extração de regras: coeficiente de suporte e coeficiente de confiança. Essas medidas são usadas para eliminar as regras que não aparecem com tanta frequência e as regras que não têm força estatística. (AGRAWAL; SRIKANT, 1994). O suporte é definido como a porcentagem de registros da base de dados que apresentam a regra de associação em questão e a confiança indica a relevância da regra com relação ao seu antecedente e quanto mais próximo de 1, mais interessante se torna essa regra.

A mineração de regras de associação foi proposta inicialmente por (AGRAWAL; FALOUTSOS; SWAMI, 1993) e baseia-se em considerar a base de dados como um conjunto de itens que forma uma coleção de transações. E uma regra de associação é uma implicação do tipo *antecedente* $\rightarrow$ *consequente* em que o antecedente e o consequente são partes distintas que compõem a regra.

Sendo $I = (i_1, i_2, ..., i_n)$ um conjunto de itens de uma base de dados, uma regra de associação é uma implicação $X \rightarrow Y$, onde $X \subset I, Y \subset I, X \cap Y = \emptyset$. Sendo X e Y, o antecedente e o consequente, respectivamente (ELMASRI; NAVATHE, 2006).

O suporte de uma regra é definido na equação 2.8:

$$Suporte(X \rightarrow Y) = \frac{\text{Total de Transações com ocorrências}(X \cup Y)}{\text{Total de Transações}} \tag{2.8}$$

E a confiança é dada pela equação 2.9:

$$Confiança(X \rightarrow Y) = \frac{\text{Total de Transações com ocorrências}(X \cup Y)}{\text{Total de Transações com ocorrências de X}} \quad (2.9)$$

O processo de extração de conhecimento por regras de associação em dados complexos apresenta uma dificuldade maior que o processo tradicional (FERREIRA et al., 2010). Essa afirmação pode ser estendida às séries temporais, pois, primeiramente, é necessário encontrar objetos similares, além disso, as regras de associação basicamente, contam a frequência de determinado objeto na base. No entanto, ao se trabalhar com séries, mesmo após a aplicação de um descritor, ela geralmente não apresentará dados iguais, o que gera uma grande dificuldade para conseguir elencar os objetos similares. Um modo de contornar essa situação é agrupando os objetos de maneira que se possa incluí-los em uma determinada classe para assim obter a frequência de um determinado conjunto e inferir regras de associação entre eles. Essa técnica é conhecida como discretização dos dados. A primeira solução foi proposta por Srikant e Agrawal (1996) em que os dados são analisados em intervalos de valores numéricos, ficando em faixas de valores numéricos ou em valores categóricos que representam o intervalo. No entanto, essa técnica apresenta uma discretização muito lenta e o número de regras geradas tende a crescer rapidamente (RIBEIRO, 2008).

2.8 Considerações finais

Neste capítulo foram apresentados os principais conceitos necessários para a análise de séries, desde as características intrínsecas e inerentes à série, até as principiais técnicas utilizadas para consulta por similaridade por meio da redução de dimensionalidade e do cálculo de distância. Além disso, os principais descritores utilizados em séries foram expostos e as técnicas mais utilizadas para a validação dos mesmos foram descritas. Com este estudo, tem-se o ferramental utilizado para a busca de similaridade em séries e para a mineração dos dados com a busca por associações que é utilizado como base desta monografia.

Capítulo 3

TRABALHO DESENVOLVIDO

Este capítulo apresenta o trabalho desenvolvido e os resultados obtidos durante a execução do presente projeto de mestrado. O capítulo está organizado da seguinte maneira: a Seção 3.1 apresenta as considerações iniciais do capítulo e a justificativa para a sua execução; na Seção 3.2, é apresentado o trabalho desenvolvido para consultas por similaridade em séries unidimensionais; na Seção 3.3 é apresentado o trabalho desenvolvido para séries multidimensionais; a Seção 3.4 apresenta o trabalho desenvolvido com consultas visuais para séries temporais. Na Seção 3.5 é exibido o método desenvolvido de mineração de regras de associação; e, na Seção 3.6, são feitas as considerações finais do capítulo.

3.1 Considerações Iniciais e Justificativas

Haja vista a evolução da computação impelida pelo aumento no poder de processamento, pela facilidade de armazenamento contínuo de grandes quantidades de dados a um custo baixo e pela introdução de novas tecnologias de captação de informações, especificamente, pelos sensores, nota-se um aumento considerável no volume de dados armazenados e manipulados pela maioria das organizações. No entanto, essa velocidade de coleta de informações é muito maior do que a velocidade de processamento, análise, síntese ou extração de conhecimento a partir desses dados coletados. Logo, torna-se necessário que sejam feitas análises sobre essa grande quantidade de dados, para que sejam estabelecidos indicadores para uma possível descoberta de padrões implícitos nos dados, assim como possíveis relações de causa e efeito, auxiliando o usuário na tomada de decisão.

Além disso, existem lacunas a serem preenchidas no campo de estudo sobre séries temporais, seja pela grande quantidade de dados que dificulta, primeiramente, o armazenamento dessas informações, como também a manipulação dos dados que exige um custo elevado de

processamento.

Outra lacuna encontra-se na redução de dimensionalidade de séries, pois as técnicas existentes apresentam certas deficiências para reduzir a dimensão e representar os dados de maneira que diminua a perda de informação e, ainda, consiga processá-los rapidamente. Assim como, a validação dos descritores para verificação se são eficazes, pois não está consolidada na literatura. Outro fato importante é que a maioria das técnicas para mineração de regras de associação para dados contínuos apresentam restrições, pois os dados precisam ser discretizados para serem minerados e isso pode elevar o custo computacional e gerar resultados indesejáveis.

Assim, a análise de séries temporais, por meio da redução de dimensionalidade, busca similaridade em séries e mineração por regras de associação, e apresenta um vasto campo de pesquisa e um grande desafio para a sua execução, motivando o trabalho deste projeto.

3.2 Séries unidimensionais

A referente pesquisa, inicialmente, focou em encontrar um descritor que consiga representar as características da série temporal e gerar descrições que contenham informações suficientes para suportar consultas por similaridade e que possibilite reduzir a dimensionalidade dos dados sem grande perda de informação. Várias técnicas são encontradas na literatura, mas não encontrou-se nenhuma que atendesse aos requisitos necessários (custo computacional baixo e ótima representatividade da série original). Assim, foi proposto um descritor baseado na lei de Coulomb (PARIS, 1788) para atender estes objetivos.

A tabela 3.1, elenca um comparativo de vantagens e desvantagens existentes nos principais descritores encontrados na literatura e com isso, justifica a elaboração de um novo descritor, haja vista, as deficiências encontradas em cada um deles.

3.2.1 Descritor baseado na Lei de Coulomb

A Lei de Coulomb estabelece a relação matemática entre a carga de dois ou mais corpos e sua força elétrica produzida, calculando as forças de interação (atração e repulsão) existentes nessas cargas. Os princípios da lei de Coulomb podem ser expressos por:

- A intensidade da força elétrica é diretamente proporcional ao produto das cargas elétricas; e

Tabela 3.1: Comparativo entre os descritores encontrados na literatura

Descritor	**Vantagens**	**Desvantagens**
Sequential Scan	apresenta boa acurácia	tem elevado custo computacional
Transformada Discreta de *Fourier*	eficiente para análise de séries com grande variação entre os dados	ruim para análise de séries estacionárias
Decomposição em valores singulares	não apresenta grande perda da representatividade da série	apresenta perda na representação de grandes subsequências e tem alto custo computacional
Transformada discreta de *wavelet*	não há	ineficiente para séries com grandes amplitudes
Piecewise aggregate approximation - PAA	custo computacional médio	apresenta perdas graves na representatividade de séries com grandes variabilidade dos dados
Adaptive Piecewise Constant Approximation - APCA	resultados melhores que o descritor PAA	custo computacional alto para séries com grande variabilidade dos dados
Piecewise Linear Aproximation - PLA	apresenta resultados satisfatórios para séries estacionárias	baixa acurácia para séries com variabilidade nos dados
Symbolic Aggregate Approximation - SAX	baixo custo computacional	perda de representatividade da série
Dynamic time warping - DTW	boa acurácia	custo computacional quadrático

- A intensidade da força elétrica é inversamente proporcional ao quadrado da distância entre os corpos.

A fórmula da lei é expressa em 3.1 :

$$\vec{F} = K\frac{q_1 q_2}{r^2}\hat{r} \tag{3.1}$$

Em que:

$\vec{F}$ é a força em Newtons;

r é a distância entre as duas cargas pontuais;

q_1 e q_2 são as intensidades das cargas;

$\hat{r}$ é o vetor unitário de direção; e

K é a constante de Coulomb.

Diante do exposto, a proposta para busca de similaridade em séries, considera as observações da série temporal como cargas puntiformes com valores de carga q constantes localizadas no plano de coordenadas formadas pelo índice da série e pelo valor da observação.

Como é necessário o cálculo da distância existente entre as cargas para obter a interação entre elas, considera-se um plano cartesiano formado pelo índice da série temporal (eixo das abscissas) e pelo valor das observações (eixo das ordenadas) e assim é possível calcular a distância entre as cargas para o cálculo das forças.

Além disso, uma carga fictícia puntiforme q é inserida no centroide composto pelos conjuntos das observações que compõem as subsequências de busca e essa carga tem por objetivo proporcionar a representação do intervalo, pois, além de se localizar no centro geométrico da subsequência, ela é utilizada para o cálculo da interação entre ela e as demais cargas gerando a força resultante que representa a subsequência.

Como a força resultante é uma medida vetorial, logo, a direção e sentido da carga influenciam o cálculo, para isso, foi estabelecido que cargas que se encontram abaixo da carga existente no centroide, possuem direção contrária àquelas que se encontram acima dela e, por consequência, apresentam intensidade negativa de força.

Dessa forma, é possível representar a série temporal por meio de um sistema de interação de partículas eletricamente carregadas e calcular a força resultante F, obtida por meio de uma soma vetorial de todas as forças que integram o sistema e assim conseguir reduzir a dimensionalidade da série para auxiliar a busca por similaridade sem grande perda de informação.

O descritor Coulomb é formalmente definido a seguir. Seja uma série temporal unidimensional $Y = (x_1, x_2..., x_n)$. Para um intervalo de interesse $P[i, j] \mid 1 \leq i \leq j \leq n$, tem-se a subsequência $S_p = (x_i, x_{i+1}..., x_j)$ correspondente ao intervalo.

O centroide de S_p é dado por $C_p = (P_p, H_p)$, onde P_p é o centro do índice das observações de S_p:

$$P_p = \frac{j-i}{2} \tag{3.2}$$

e, H_p é a altura média da subsequência S_p e é dada pela média das medidas de S_p, dada por:

$$H_p = \frac{\sum_{b=i}^{j} x_b}{|S_p|}, \tag{3.3}$$

onde $|S_p|$ é o número de observações da subsequência S_p.

O módulo da distância de uma observação $Q_p \in S_p$ até o centroide é dado por:

$$r = \sqrt{\sum_{k=1}^{n} (|C_p - Q_p|)^2} \tag{3.4}$$

E a direção vetorial $\hat{r}$ é dada por: $\hat{r} = \vec{C_p} + \vec{Q_p}$

A carga puntiforme $q_(s_p)$ em S_p é dada por $q(x_b) = -q$, se $x_b < H_p$; 0 se $x_b = H_p$; $+q$ se $x_b > H_p$ e a carga puntiforme em $q(c_p) = q$.

A força resultante $\vec{F}$ das cargas de S_p sobre a carga colocada em C_p é dada por:

$$\vec{F} = \sum_{a=i}^{j} \frac{q(c_p).q(s_a)}{r^2} \hat{r}, \tag{3.5}$$

Assim, a subsequência S_p é representada por $S_p \rightarrow < \vec{F_p}, H_p >$.

Nesta abordagem proposta, o vetor de características é formado pela força resultante calculada na subsequência de interesse e, também, pela altura do centroide. Conforme exibido na Expressão 3.6:

$$V = [\vec{F}, H] \tag{3.6}$$

A necessidade do uso da altura do centroide se justifica, pois a força resultante consegue mapear a interação entre os pontos do intervalo que a compõe, no entanto, nenhuma informação com relação à altura existente entre os dados originais é armazenada e essa informação é importante para o cálculo da similaridade.

Para definir o grau de semelhança entre as instâncias, utilizando o vetor de características anteriormente descrito, é aplicada a distância Euclidiana. O uso desta distância se justifica, pois

ela se aplica melhor aos dados e à fórmula da Lei de Coulomb, mantendo a intensidade da força elétrica inversamente proporcional ao quadrado da distância entre os corpos.

Na Figura 3.1 é ilustrada a representação gráfica de uma subsequência da série com a inserção de cargas puntiformes nos valores das observações e de uma carga neutra inserida no centroide do intervalo e na Figura 3.2 ilustra-se a convenção da intensidade das cargas adotadas. A Figura 3.3 mostra a interação entre a carga inserida no centroide e as demais cargas existentes no intervalo.

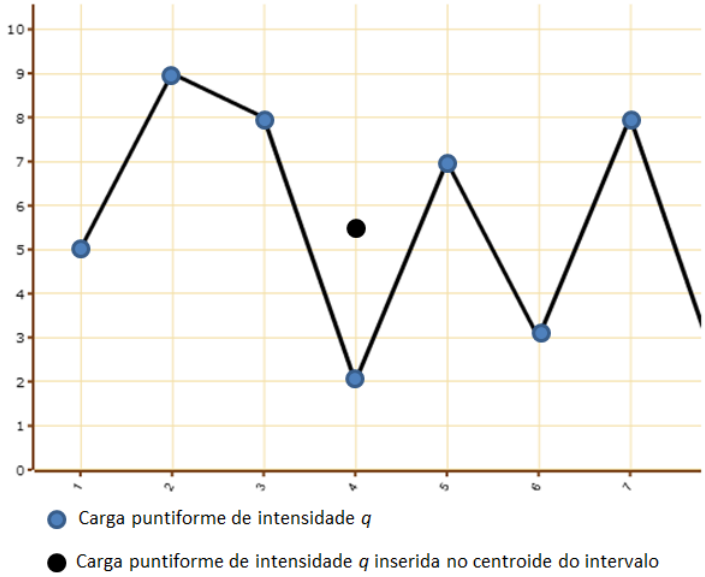

Figura 3.1: Representação de uma subsequência da série formada por 7 observações com suas respectivas cargas puntiformes inseridas nas observações e uma carga neutra inserida no centroide.

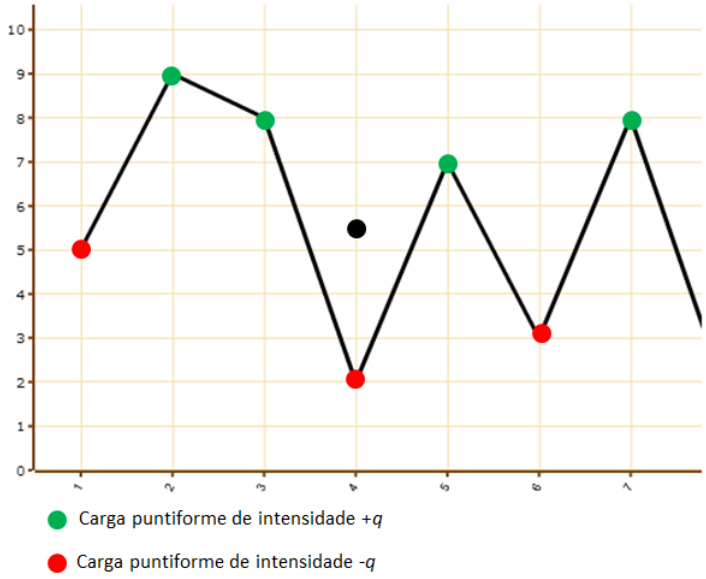

Figura 3.2: Representação da intensidade da carga com relação à carga localizada no centroide.

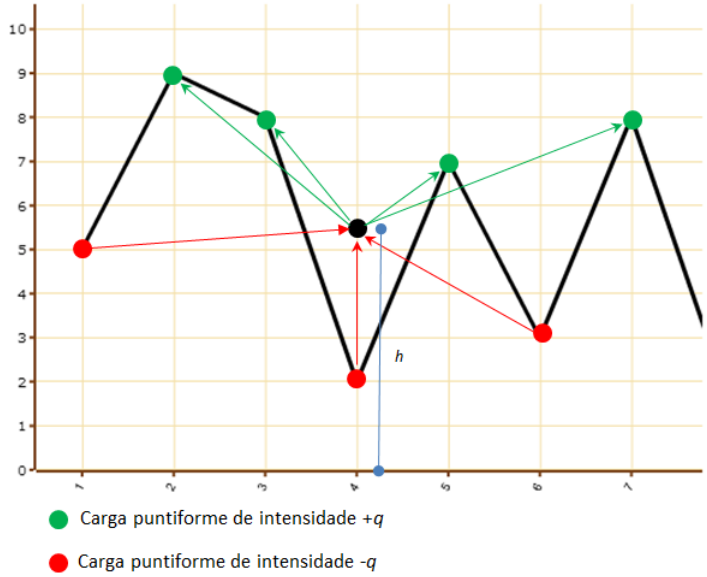

Figura 3.3: Interação entre as cargas da série e a carga inserida no centroide.

Na Figura 3.4 é exibido o algoritmo principal do descritor Coulomb e na figura 3.5 uma função auxiliar para o cálculo da força.

O algoritmo do descritor Coulomb recebe como entrada uma série temporal e uma subsequência de interesse definida pelo usuário e retorna as subsequências ordenadas pelo grau de similaridade fazendo uma consulta do tipo *knn* com k igual à quantidade de intervalos que cabem na série sem repetição.

Na *linha 1* o algoritmo inicia a leitura dos dados procurando por subsequências relevantes, ou seja, prováveis candidatos similares ao objeto de interesse por meio da análise de pontos crescentes ou decrescentes. Nas *linhas 2, 3 e 4*, o algoritmo calcula a força resultante de interação entre as cargas por meio do algoritmo auxiliar exibido na Figura 3.5. Após isso, há a ordenação do vetor que contém os valores da força resultante de acordo com a similaridade entre esses intervalos e o intervalo de interesse (*linhas 5, 6 e 7*) e o resultado é exibido para o usuário na *linha 8*.

O algoritmo auxiliar para o cálculo da força recebe como entrada uma subsequência da série e retorna um vetor com a força resultante e a altura do centroide. Na *linha 1*, o algoritmo calcula o centroide do intervalo. E para cada ponto pertencente à subsequência é calculada a interação entre esse ponto e o centroide pela Lei de Coulomb (*linhas 2 e 3*). A força resultante é obtida pela soma das forças calculadas em cada ponto e o resultado é retornado para o algoritmo principal (*linhas 4 e 5*).

Algoritmo: *Coulomb*

Entrada:

- Uma série temporal Y na forma $(x_1, x_2, ..., x_n)$;
- Uma subsequência de interesse

Saída: subsequências da série ordenadas pelo grau de similaridade.

1. percorra a base de dados
2. **para cada** intervalo relevante da série **faça**
3. $vetor[][]$ = Calcule $\vec{F}(intervalo)$
4. $x[][]$ = Calcule $\vec{F}(interesse)$
5. **ordene** $vetor[]$ de acordo com a proximidade com x
6. **para cada** valor $\vec{F}$ de $vetor$ **faça**
7. $result[][] = [F, altura]$
8. **escreva** $result$

Figura 3.4: Algoritmo para cálculo da força resultante F.

Algoritmo: *Cálculo de* $\vec{F}$

Entrada:

- Subsequência da série Y na forma $(x_1, x_2, ..., x_n)$;

Saída: Força resultante $[F, h]$.

1. $C = centroide(F)$
2. **para cada** ponto do intervalo da série Y **faça**
3. $forca = Coulomb(C, P)$
4. $F = \sum_{a_1}^{a_n} forca$
5. **retorne** $[F, altura(C)]$

Figura 3.5: Algoritmo auxiliar para cálculo da força resultante.

3.2.2 Resumo do descritor Coulomb

Em suma, o descritor Coulomb apresenta-se como um potencial descritor para reduzir a dimensionalidade das séries temporais. Isso ocorre devido ao fato de que dada uma subsequência da série $Y = (x_1, x_2, ..., x_n)$ em que n é o tamanho dela, o descritor Coulomb consegue representá-la através de duas medidas [*F, H*] em que:

- F : representa a força existente na interação entre as observações da subsequência. Ela é adequada para representar essa interação, pois é proporcional à distância que separa as cargas com relação a carga colocada no centroide e ela consegue prover a tendência da subsequência; e
- H : é a altura do centroide. Ela representa a média dos valores da subsequência e é adequada para representar o comportamento geral dos dados.

Dessa forma, o vetor de características formado por F e H tende a representar a série temporal de maneira apropriada. Proporcionando, assim, o uso do descritor Coulomb para a manipulação de séries temporais e seu uso em consultas por similaridade.

3.2.3 Experimentos e Resultados Obtidos

Com o intuito de validar o método proposto, foi desenvolvido um protótipo que realiza consultas do tipo k-vizinhos mais próximos (kNN). O protótipo realiza as seguintes tarefas: i) quando inserida uma série, ele gera uma visualização gráfica dos dados para que o especialista indique qual intervalo é interessante para análise; ii) após escolhida a subsequência de análise e o valor de k, é gerado um novo gráfico onde há o hachuramento dos intervalos do gráfico que apresentam maior similaridade com o objeto definido. E também uma tabela exibindo as subsequências em ordem crescente de similaridade.

A Figura 3.6 ilustra um exemplo de consulta feita usando o protótipo desenvolvido.

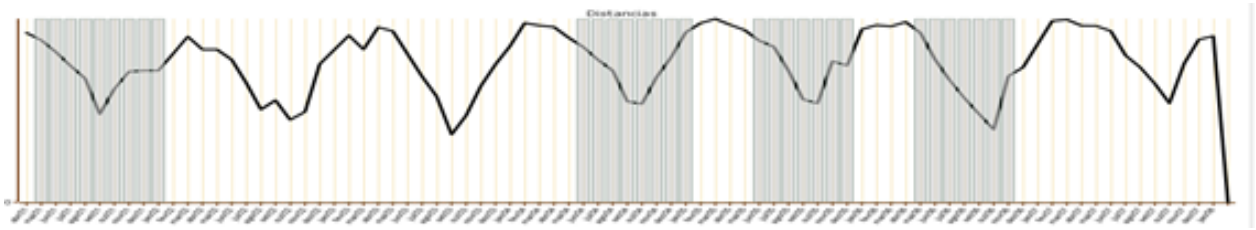

Figura 3.6: Gráfico exibindo o resultado de uma consulta com *knn* = 3.

Para a realização dos experimentos foram utilizados dados meteorológicos obtidos do projeto Agrodatamine (AGRODATAMINE..., 2013) em que há sensores de medições de dados meteorológicos como temperatura, precipitação pluviométrica, umidade relativa do ar de várias cidades brasileiras obtidos diariamente com medidas que se iniciam no ano de 1950 até os dias atuais. Também foram utilizados dados obtidos em (CLIMATE-PREDICTION-CENTER, 2012) relativos à temperatura média da superfície do mar na região 3.4 de El niño, dados de bases de dados aleatoriamente gerados para experimentos e dados de uma base de dados médica obtida em UCI *Machine Learning Repository* (UCI..., 2013) em que há dados de nível de glicose de pacientes no decorrer de atividades diárias.

O descritor proposto foi comparado com o método *Sequential Matching* (SM), com o descritor *Discrete Fourier Transform* (DFT) e também, com o *Dynamic time warping* - DTW, pois estes métodos são considerados *baselines* do trabalho em questão. O primeiro por apresentar uma acurácia alta, o segundo por ter um bom desempenho para grandes bases de dados e o terceiro por ser amplamente utilizado pela comunidade científica e por apresentar boa acurácia.

Acurácia

Como experimento inicial para verificar a acurácia do método proposto foram utilizadas amostras da base de dados Agrodatamine, em que os dados de temperatura mínima de uma cidade brasileira (Alegre, ES) foram obtidos por meio de amostras mensais do ano de 1979 a 2010 e foram utilizados para localizar as subsequências de maior similaridade de acordo com uma determinada estação do ano. No caso deste experimento, foram consultados as 10 subsequências mais similares (*knn-query*) ao período referente ao inverno brasileiro (de 21 de junho a 23 de setembro) de 1979. O resultado da consulta é exibido na figura 3.7, as subsequências mais similares estão hachuradas.

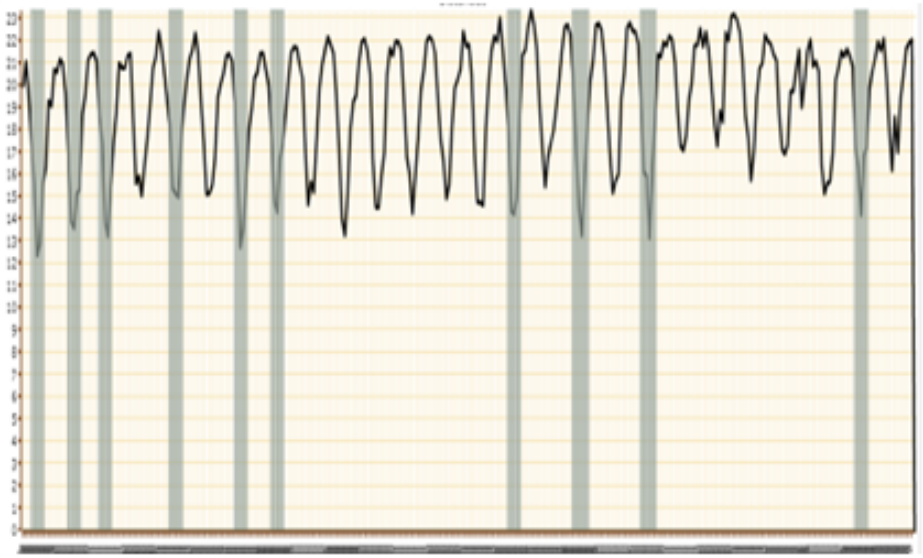

Figura 3.7: *knn-query* com *n* = 10 para dados relativos ao inverno brasileiro de 1979.

Da realização dos experimentos, foi possível comprovar que a consulta retorna os elementos que apresentam os períodos de inverno com temperatura média próxima da temperatura do objeto de consulta.

Outro experimento para verificar a acurácia do método proposto foi realizado utilizando a base de dados meteorológica de Climate... (2013). Onde foram realizadas consultas para localizar as subsequências de maior similaridade de acordo com uma determinada estação do ano. No caso deste experimento, foram realizadas 10 consultas procurando localizar as 100 subsequências mais similares (*knn-query*) à subsequência informada pelo usuário. Por exemplo, foi realizada uma consulta visando encontrar as subsequências da série referente ao verão (de 21 de junho a 23 de setembro) e ao inverno norte-americano (de 21 de dezembro a 20 de março do ano seguinte) do ano de 1900 e as subsequências positivas retornadas pela consulta foram contabilizadas. As consultas foram realizadas utilizando os descritores em análise e o resultado

da acurácia média é exibido na Tabela 3.2.

Pela execução de consultas por similaridade em séries unidimensionais, nota-se que os resultados apresentados pelo descritor Coulomb são satisfatórios para a consulta por similaridade.

Tabela 3.2: Comparativo de acurácia entre os descritores em análise

	DFT	SM	DTW	Coulomb
Acurácia	20,48%	46,63%	66,5%	**68,95%**

Complexidade Computacional

Para a verificação da complexidade do algoritmo foram realizados experimentos utilizando-se bases geradas aleatoriamente com o intuito de verificar o desempenho do descritor Coulomb, comparando-o primeiramente com o método *Sequential Matching* (SM), com o *Discrete Fourier Transform* (DFT), e também, com o *Dynamic time warping* - (DTW). Os experimentos foram executados em um computador com processador *Intel*(R) *Core*(TM) i7-860 de 2,8GHz com 8,00 GB de memória RAM e sistema operacional *Microsoft Windows 7* de 64 bits.

O primeiro experimento para a verificação da complexidade dos algoritmos consistiu em executar uma mesma consulta *knn* utilizando-se os três descritores, variando o tamanho da base de dados e registrando o tempo gasto para a execução das consultas. Pois, conforme aumenta o tamanho da base a quantidade de cálculos executados também aumenta. O gráfico da Figura 3.8 mostra os tempo de consulta para diferentes tamanhos de base de dados.

Tabela 3.3: Tempo em segundos para a execução de uma consulta por similaridade realizada pelos 4 descritores em análise variando o tamanho da base

Tamanho	**SM**	**DFT**	**DTW**	**Coulomb**
100	0,050	0,097	0,043	0,037
1000	0,951	0,269	0,309	0,034
2000	4,072	0,680	0,455	0,050
3000	12,818	2,255	0,475	0,398
4000	22,683	2,262	0,848	0,744
5000	49,210	7,863	0,959	0,082
6000	73,226	7,773	1,721	1,388
7000	91,616	8,856	2,453	1,933
8000	161,927	7,818	2,647	0,588
9000	238,786	39,399	2,848	3,750
10000	319,583	32,006	3,810	3,746

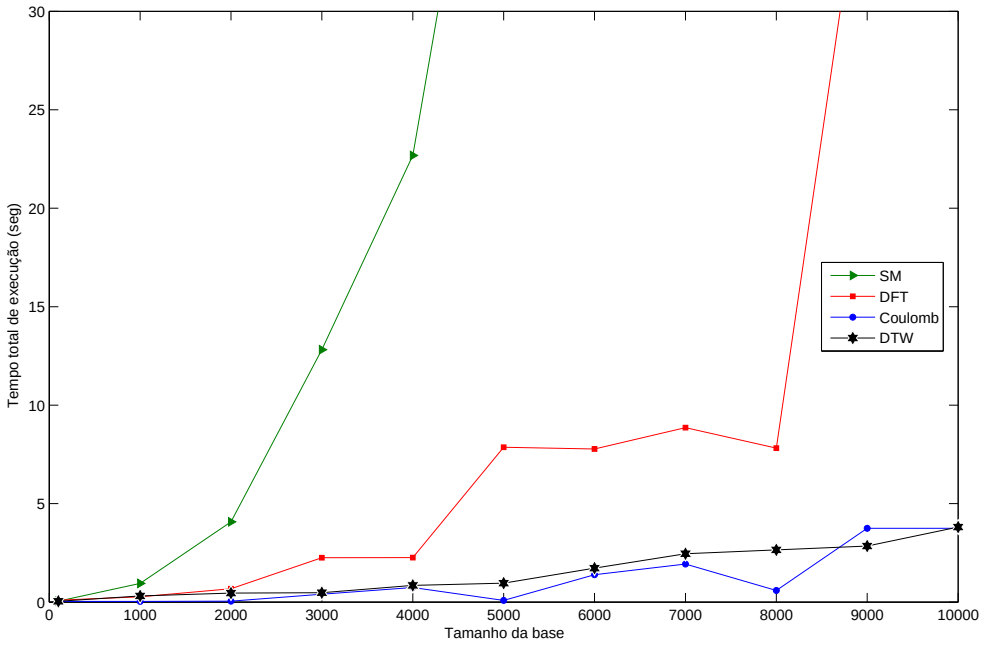

Figura 3.8: Tempo gasto por consulta variando o tamanho da base de dados (bases de dados gerados aleatoriamente).

Conforme demonstra o gráfico da Figura 3.8 e a tabela 3.3, o descritor Coulomb apresenta um tempo de execução menor que os descritores SM e DFT e bem próximo ao descritor DTW. Independentemente do tamanho da base, o descritor Coulomb busca somente subsequências que são possíveis candidatas a serem similares com a subsequência de interesse. Já os descritores DFT e SM fazem uma varredura completa dos dados e executam cálculos para subsequências que posteriormente serão desprezadas. Assim como o descritor DTW busca o caminho mínimo entre todas as subsequências existentes na série temporal.

Outro experimento realizado para verificar a eficiência com relação à complexidade foi uma consulta *knn* em uma base de dados sintética e variar o tamanho da subsequência consultada para observar o comportamento dos métodos com relação ao tempo gasto para a execução da consulta. A Figura 3.9 mostra o gráfico com essas medidas para os métodos e a tabela 3.4 exemplifica algumas medidas pontuais.

Pela análise do gráfico da Figura 3.9 e da tabela 3.4, nota-se que o descritor Coulomb apresenta um tempo inferior ao método SM e se comparado ao descritor DTF, ele apresenta bons resultados para subsequências de consulta inferiores a 400 dados. Com relação ao descritor DTW os resultados do descritor Coulomb ficam muito abaixo devido ao fato de que o DTW apresenta uma complexidade quadrática em seus cálculos. Para a análise de consultas em bases meteorológicas, as subsequências, na prática, não são maiores que um ano, ou seja, 365 dados,

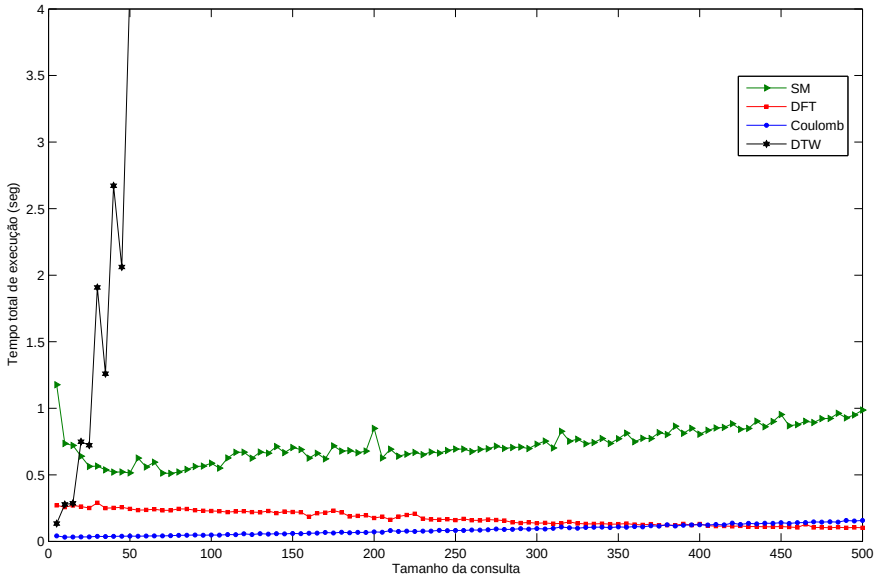

Figura 3.9: Tempo gasto por consulta variando o tamanho da consulta.

Tabela 3.4: Tempo em segundos para a consulta por similaridade variando o tamanho da subsequência de consulta.

Tamanho do intervalo	SM	DFT	DTW	Coulomb
10	0 ,159	0 ,236	0 ,278	0 ,032
20	0 ,289	0 ,234	0 ,749	0 ,034
30	0 ,269	0 ,237	1 ,908	0 ,038
40	0 ,313	0 ,249	2 ,672	0 ,038
50	0 ,262	0 ,238	4 ,173	0 ,039
60	0 ,297	0 ,244	5 ,509	0 ,041
70	0 ,329	0 ,153	7 ,611	0 ,042
80	0 ,308	0 ,163	10 ,308	0 ,046
90	0 ,381	0 ,154	13 ,083	0 ,048
100	0 ,311	0 ,155	13 ,806	0 ,048
150	0 ,410	0 ,154	37 ,803	0 ,059
200	0 ,621	0 ,168	57 ,027	0 ,071
250	0 ,824	0 ,160	68 ,671	0 ,082
300	1 ,138	0 ,137	91 ,289	0 ,096
350	1 ,164	0 ,130	158 ,686	0 ,109
400	1 ,192	0 ,131	188 ,309	0 ,125
450	1 ,083	0 ,111	218 ,213	0 ,141
500	1 ,318	0 ,102	262 ,176	0 ,158

logo o descritor Coulomb, também apresenta resultados satisfatórios com relação à complexidade.

Precisão x Revocação

Para a realização dos experimentos de precisão x revocação, foram utilizados dados da base Agrodatamine, do National Weather Service (KMNI. . . , 2013) referentes aos dados do El niño, uma base de dados meteorológica contendo a temperatura média mensal do Central Park em Nova Iorque-EUA e uma base médica com a quantidade de glicose no sangue de pacientes que fazem o uso de insulina. Para a elaboração dos gráficos de precisão e revocação foram utilizadas as recomendações descritas em Meadow, Boyce e Kraft (2000).

Na primeira base foram utilizados dados referentes à temperatura mínima de uma cidade brasileira e dez consultas por similaridade. Os descritores anteriormente citados foram executados buscando estações do ano similares ou períodos em que há quedas ou aumentos de temperatura fora do padrão normal. Dos dados obtidos foram extraídos a precisão e revocação para cada ponto de interesse e o gráfico comparativo foi elaborado e é apresentado na Figura 3.10.

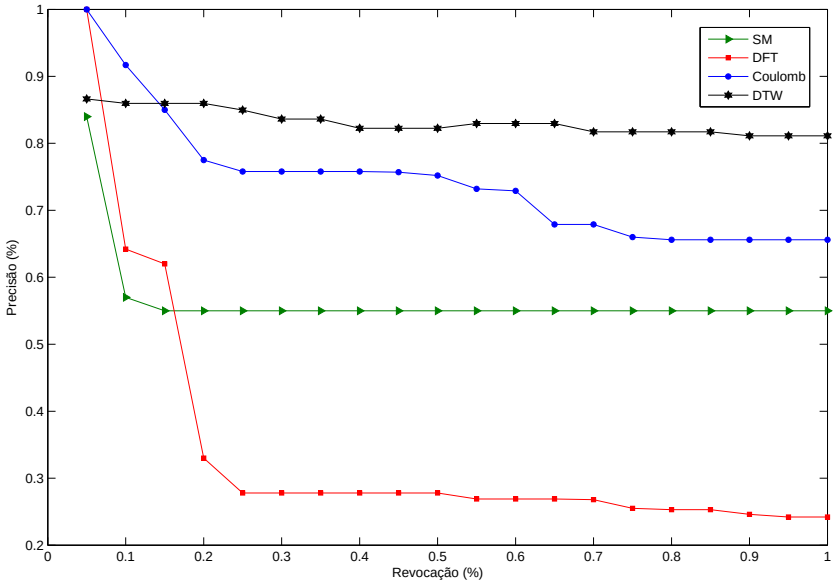

Figura 3.10: Precisão x revocação para a base de dados *Agrodatamine*.

Outro experimento realizado utilizou a temperatura da superfície do oceano na região 3.4 (SST in the *Niño*-3.4 region) onde ocorre o fenômeno do *El niño*. Tendo em vista que esse fenômeno é cíclico e tem um ciclo a cada 12 anos, em que a média da temperatura no ciclo é maior com o passar dos anos, foram realizadas consultas por similaridade buscando intervalos de anos ou de meses que pertencem ao mesmo ciclo.

Os experimentos foram realizados com os descritores em análise e foi construído o gráfico

de precisão e revocação apresentado na figura 3.11.

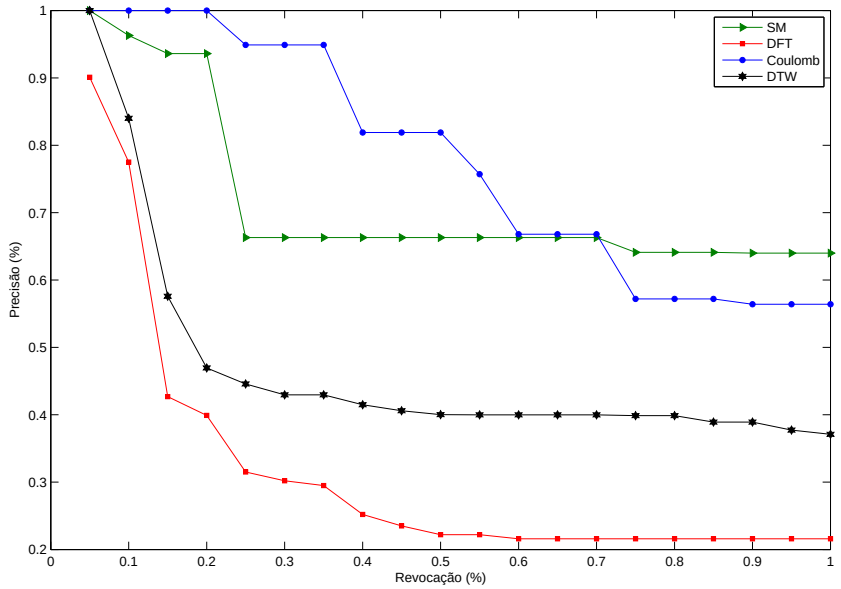

Figura 3.11: Precisão x revocação para a base de dados SST na região 3.4 de *El niño*.

Examinando o gráfico da Figura 3.11, nota-se que o descritor Coulomb tem resultados satisfatórios para a base de dados em análise, pois ele apresenta uma bom índice de precisão e revocação quando comparado com os demais métodos.

Na base meteorológica do Central Park foram utilizados dados referentes à temperatura média mensal da cidade de Nova York. Os descritores comparados foram executados buscando estações do ano similares, períodos em que há quedas ou aumentos de temperatura fora do padrão normal e períodos com alguma variabilidade cíclica existente na temperatura. Dos dados obtidos foram extraídos a precisão e revocação para cada ponto de interesse e o gráfico comparativo foi elaborado e é apresentado na Figura 3.12.

Outro experimento realizado utilizou a base médica. Tendo em vista que o nível de glicose de um paciente diminui após a aplicação de insulina, os experimentos se basearam em buscar períodos de alto ou baixo nível de glicose no sangue de pacientes antes ou após a aplicação de insulina no organismo, e também, em períodos específicos do dia como: antes ou após às refeições ou de manhã ou à noite. Os experimentos foram realizados e foi construído o gráfico de precisão e revocação apresentado na Figura 3.13.

Pela análise dos gráficos de precisão versus revocação elaborados, nota-se que o descritor Coulomb apresenta um resultado bastante satisfatório com relação aos descritores SM e DFT. A precisão, no geral, é alta para uma revocação menor que 50% enquanto os descritores SM e

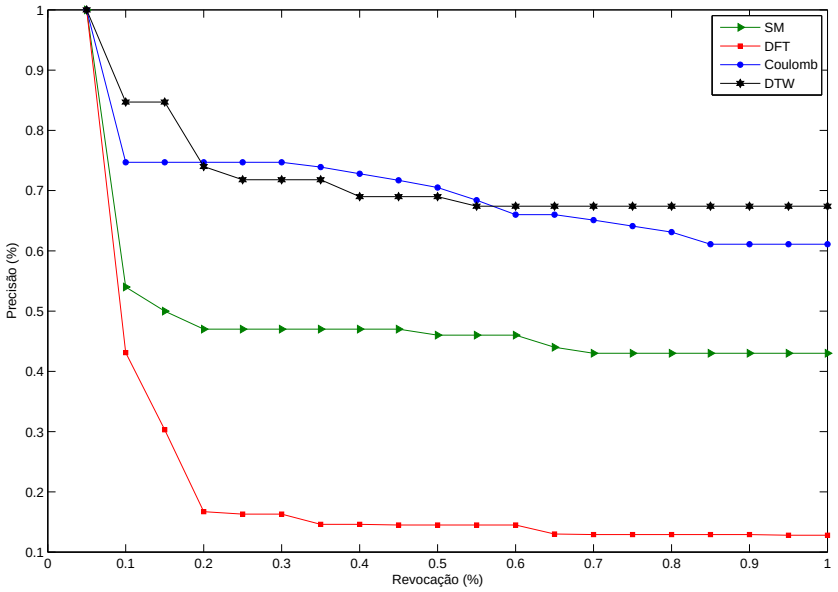

Figura 3.12: Precisão x Revocação para a base do Central Park.

DFT apresentaram baixa precisão para índices de revocação também baixos.

Com relação ao descritor DTW, que em alguns casos, apresentou resultados bem próximos ou melhores que o descritor Coulomb, nota-se, pelos testes, que em séries temporais em que os dados apresentam menor variabilidade, o seu desempenho é melhor. No entanto, em séries temporais em que há grande variabilidade dos dados ou que as séries são não estacionárias o descritor Coulomb apresenta melhores resultados. Isso ocorre devido ao fato de que o descritor DTW tem bom desempenho para encontrar subsequências deslocadas no tempo (BERNDT; CLIFFORD, 1994), mas para dados com amplitude variada pode ser que ele não consiga encontrar o melhor e/ou menor caminho entre as subsequências em análise. Esse fato, no entanto, não tem interferência no descritor Coulomb.

Conforme exposto, o descritor Coulomb proposto atende às necessidades iniciais do projeto de realizar buscas por similaridade em séries temporais. Além disso, se comparado aos principais métodos encontrados na literatura, ele apresenta um ganho significativo de redução de dimensionalidade das séries, aumento da acurácia e diminuição do tempo de consulta.

3.3 Séries multidimensionais

Como etapa seguinte do projeto, a pesquisa focou-se em utilizar o descritor Coulomb para a busca de similaridade em séries multidimensionais conforme detalhado nas subseções seguin-

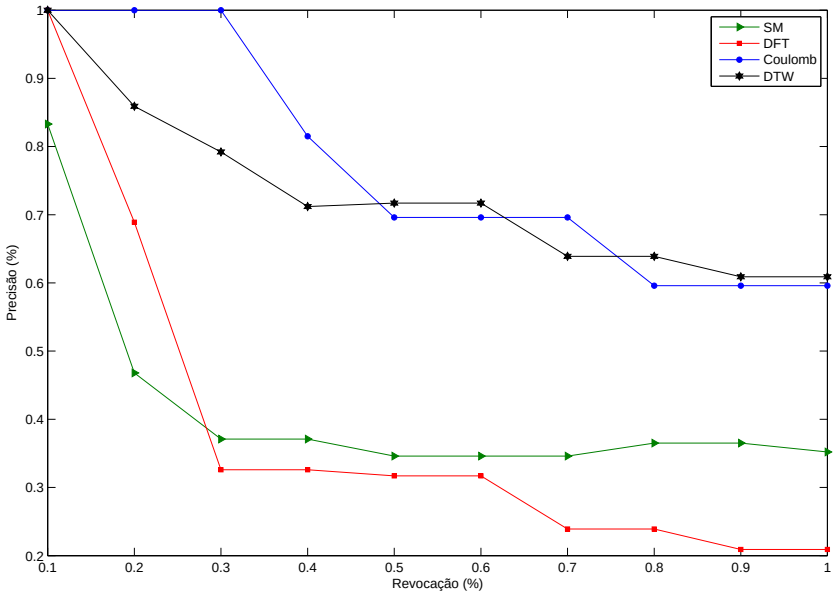

Figura 3.13: Precisão x Revocação para a base médica

tes.

3.3.1 Tractable Similarity Searching (TSS)

Para realizar a busca por similaridade em múltiplas séries temporais, um dos grandes desafios é conseguir encontrar no grupo de subsequências similares de cada uma das séries, o intervalo que apresenta as subsequências com maior similaridade ao objeto de consulta. Esse problema ocorre devido ao fato de que os intervalos similares de uma série podem não ser os mesmos para as demais séries.

Diante do exposto, foi proposto um método para a busca de similaridade em múltiplas séries, chamado de método maleável de busca por similaridade, *Tractable Similarity Searching (TSS)*. Esse método é composto por dois módulos: i) Um módulo que contém o descritor Coulomb utilizado para reduzir a dimensionalidade das séries temporais em análise e proporcionar a formação do vetor de características utilizado na busca por similaridade; e ii) um módulo chamado módulo flexível *Flexible module (FM)* para realizar a consulta por similaridade em séries multidimensionais.

O módulo flexível - FM baseia-se no princípio dos caminhos mínimos, em que, cada intervalo similar de uma série, calculado utilizando-se de um descritor, é considerado um vértice de um grafo e a ligação entre os vértices de uma série com os vértices das demais séries formam as arestas. E, além disso, os pesos das arestas são formados pelo grau de dissimilaridade existente

nas subsequências da série até um ponto comum definido previamente. Logo, há formação de um grafo $G = (V, E)$ em que o peso do caminho $p =< v_0, v_1, ..., v_k >$ é o somatório dos pesos de suas arestas constituintes: $w(p) = \sum_{k}^{i=1} w(v_i - 1, v_i)$. Dessa forma, para definir o menor caminho é necessário encontrar o menor caminho existente entre u até um ponto predefinido v dado pela equação 3.7.

$$\delta(u, v) = min\{w(p) : u \rightarrow v\} \tag{3.7}$$

Assim, tendo as distâncias dos caminhos mínimos calculadas para cada intervalo é possível elencar os trechos das séries temporais que apresentam maior similaridade de acordo com o intervalo de consulta e com isso encontrar trechos similares em séries temporais multidimensionais.

O método FM pode ser definido como: dada uma série temporal multidimensional $Y_m = (x_{11}, ..., x_{m1}), ..., (x_{1n}, ..., x_{mn})$ de tamanho n e um intervalo de interesse $P[i, j] \mid 1 \leq i \leq j \leq n$, têm-se subsequências $S \in Y_m$.

Qualquer que seja S, é possível reduzir a dimensionalidade dessa subsequência com a utilização de um descritor representando-o como uma medida de similaridade $p(S)$.

Um grafo $G = (V, E)$ é construído onde os vértices do grafo são formados pelos pontos médios das subsequências das séries de Y_m e são dados por:

$$V[S] = \frac{j - i}{2} \tag{3.8}$$

E as arestas são formadas pela ligação de $V[S_k]$ com $V[S_{k+1}]$. Os pesos das arestas são formados por $p(S_k)$ até um ponto arbitrário x com $x > m$. Assim, é possível calcular o caminho mínimo utilizando a fórmula expressa em 3.7. As figuras 3.14, 3.15, 3.16 e 3.17 ilustram o processo de consultas por similaridade em séries temporais multidimensionais:

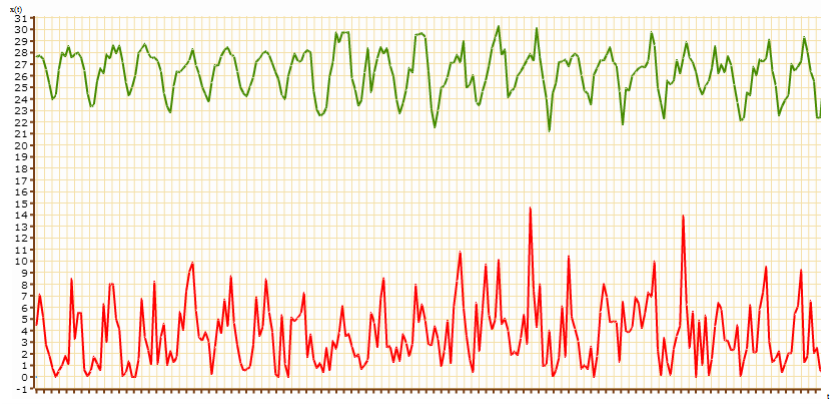

Figura 3.14: Passo 1 - São dadas as séries temporais e as subsequências de interesse.

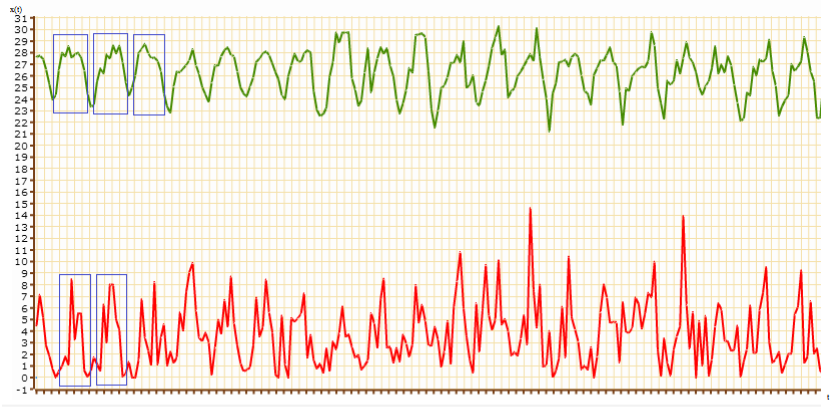

Figura 3.15: Passo 2 - O módulo com o descritor Coulomb retorna as subsequências similares conforme interesses do usuário.

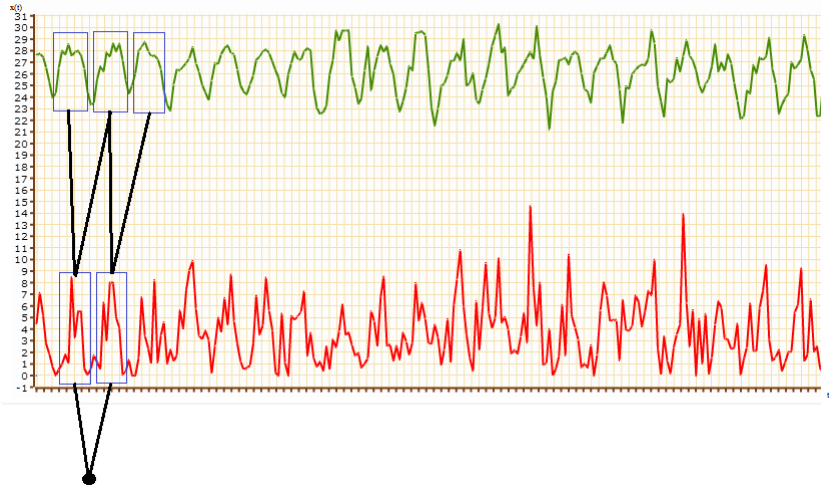

Figura 3.16: Passo 3 - O módulo *FM* projeta os caminhos existentes entre as subsequências de uma série para as demais subsequências das séries restantes.

3.3.2 Experimentos e Resultados Obtidos

Com o objetivo de validar o TSS, foram realizados experimentos para sua validação em busca por similaridade em séries temporais multidimensionais. Além disso, o módulo que contém o descritor Coulomb, foi avaliado com a utilização de outros descritores para verificar a eficácia do módulo. O método foi avaliado com relação ao seguinte aspecto:

Precisão x Revocação

Para a validação do método de busca por similaridade em múltiplas séries temporais, o método foi implementado utilizando os seguintes descritores: i) *Sequential Matching*: por apre-

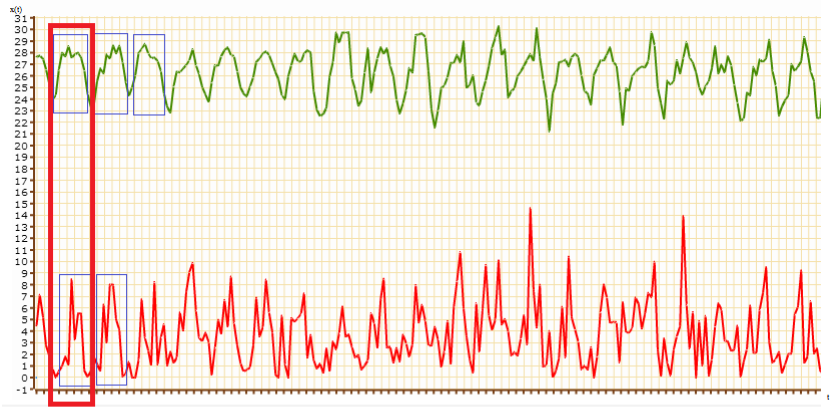

Figura 3.17: Passo 4 - O módulo *FM* calcula o menor caminho de acordo com os pesos das arestas formados pelas similaridades e retorna ao usuário o intervalo que contém o menor caminho existente entre as subsequências.

sentar uma alta acurácia nas consultas realizadas individualmente, ii) *Discrete Fourier Transform*: por apresentar uma boa performance para grande quantidade de dados e iii) Descritor Coulomb: por apresentar boa acurácia e boa performance nos experimentos realizados anteriormente. Para os três descritores foram realizados diversos experimentos e gráficos de precisão e revocação foram elaborados, utilizando as recomendações descritas em (MEADOW; BOYCE; KRAFT, 2000). Para os experimentos foram utilizadas as seguintes bases de dados:

- Séries temporais da temperatura média mensal de sete aeroportos localizados em diferentes estados dos Estados Unidos do período de 1939 a 2011, obtidos em KMNI Climate Explorer (CLIMATE..., 2013). Nessa base foram realizadas consultas sobre as estações do ano e períodos de picos de temperatura.

- Série temporal contendo a temperatura média mensal do estado da Flórida/EUA e série temporal da produtividade mensal de laranja, em toneladas, dos anos de 1983 a 2006, obtidos em KMNI Climate Explorer (KMNI..., 2013) e Climate Prediction Center (CLIMATE..., 2013). Nessas séries foram realizadas consultas sobre as estações do ano e sobre o períodos de alta, baixa e média produtividade agrícola.

- Séries temporais correspondentes a temperatura mínima, máxima e índice de precipitação mensal das cidades de Avaré e Presidente Prudente do estado de São Paulo, Brasil, obtidos em Projeto Agrodatamine (AGRODATAMINE..., 2013) dos anos de 1961 a 2010. Nessas séries foram realizadas consultas sobre o período de alta e baixa pluviosidade e períodos de alta e baixa temperatura não relacionados com a pluviosidade.

O gráfico de precisão x revocação apresentado na Figura 3.18 referente à temperatura média dos aeroportos norte-americanos, mostra que os resultados obtidos com o método TSS utilizando os três descritores supracitados. Levando-se em consideração que foram feitas consultas em sete séries bem heterogêneas, as consultas realizadas apresentaram bons índices de precisão para baixos índices de revocação e o descritor Coulomb apresentou os melhores resultados, superando até o método Sequential Matching.

A Figura 3.19 apresenta o gráfico de precisão x revocação para as séries temporais de temperatura média mensal x produção de laranja e pela análise do gráfico,nota-se que o método em análise apresentou resultados satisfatórios e que o descritor Coulomb superou os demais descritores comparados.

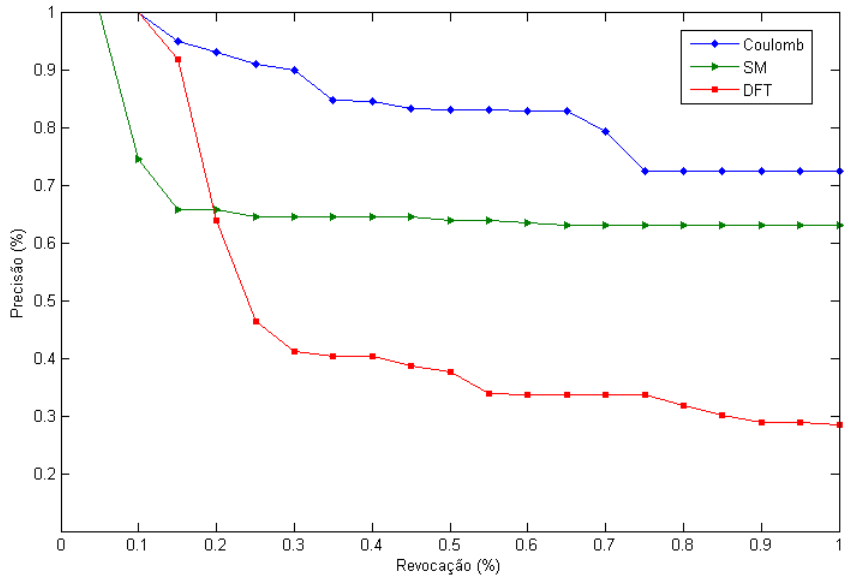

Figura 3.18: Precisão x Revocação para as séries temporais dos aeroportos.

Nas Figuras 3.20 e 3.21 são exibidos os gráficos de precisão x revocação para a temperatura mensal mínima, máxima e precipitação de duas cidades brasileiras. Pela análise do gráfico, nota-se que o comportamento do método de busca de similaridade em múltiplas séries TSS apresenta resultados satisfatórios, pois tratam-se de consultas realizadas em três séries, sendo que uma delas apresenta escala completamente diferente das demais. Além disso, o descritor Coulomb apresenta uma vantagem sobre o descritor Sequential Matching, primeiramente por ter menor complexidade e por apresentar uma acurácia melhor que o principal descritor existente na literatura.

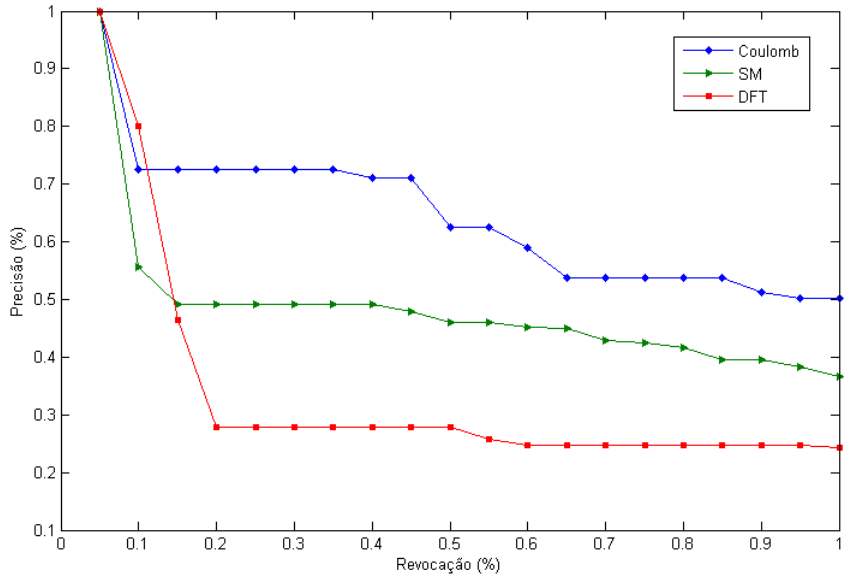

Figura 3.19: Precisão x Revocação para as séries temporais de produção de laranja.

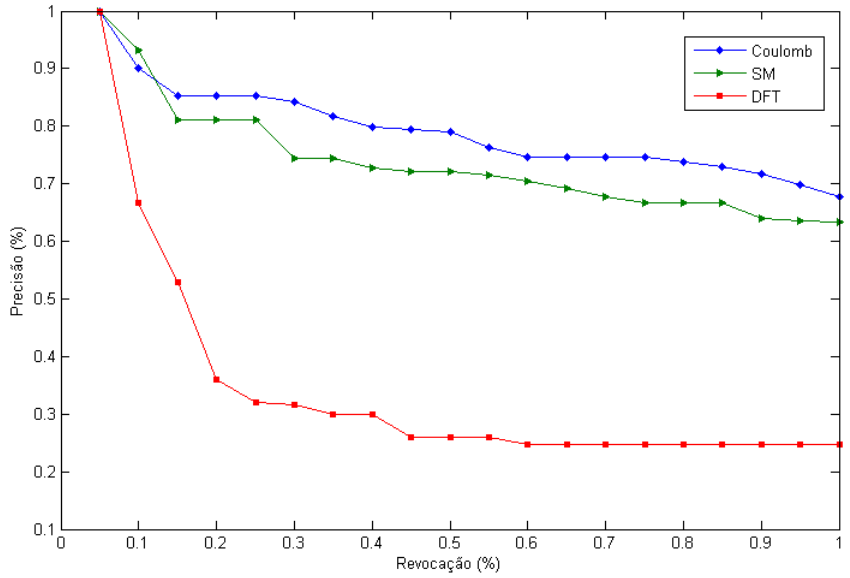

Figura 3.20: Precisão x Revocação para as séries temporais da cidade de Avaré.

3.4 Consultas Visuais por similaridade

Essa etapa do trabalho teve por objetivo propor um ambiente integrado de consultas visuais em séries temporais com a integração do descritor baseado na lei de Coulomb para a redução da dimensionalidade e um sistema de mineração visual das consultas executadas. A validação é confirmada por experimentos com dados reais de variados tamanhos e dimensões, que mostram que o sistema apresenta resultados satisfatórios para a execução de consultas visuais.

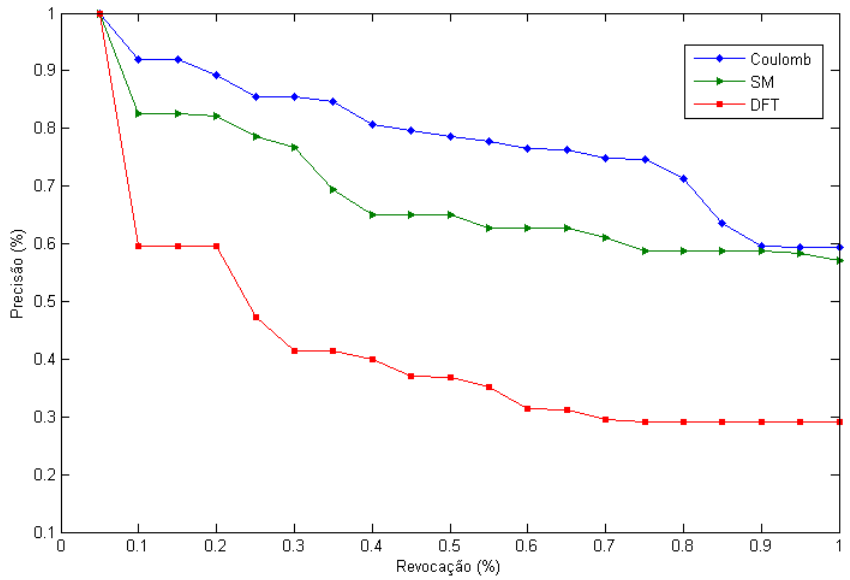

Figura 3.21: Precisão x Revocação para as séries temporais da cidade de Presidente Prudente

O sistema proposto é integrado por módulos distintos que compartilham dados entre si e trabalham harmonicamente recebendo as informações passadas pelo usuário para a realização das consultas, aplicando o descritor Coulomb aos dados de acordo com o interesse do usuário e retornando em uma resposta gráfica, com os objetos de interesse conforme encontrados e elencados pelo descritor.

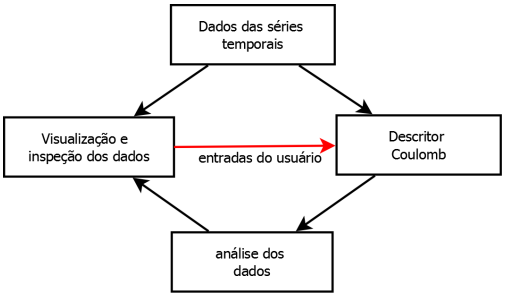

Figura 3.22: Relação existente entre os módulos.

A Figura 3.22 ilustra a relação existente entre os módulos. Os dados das séries temporais servem de entrada para o módulo de visualização e exploração dos dados onde o especialista pode verificar o comportamento e as características relevantes das séries e selecionar os intervalos interessantes para análise. E, também, servem de entrada para o descritor Coulomb que, por meio da redução de dimensionalidade e cálculo da similaridade, passa ao módulo de análise dos dados aqueles intervalos com algum grau de similaridade segundo o interesse do usuário. A partir daí, o módulo de análise dos dados elabora informações que são passadas novamente para o módulo de visualização e exploração dos dados que, por sua vez, os mostra ao usuário.

O módulo de visualização e exploração dos dados (VDEM) é responsável por toda a interação existente entre o sistema e o usuário. É por meio dele que o usuário insere, primeiramente, as séries temporais ou a série temporal multidimensional e o módulo gera uma visualização gráfica ao usuário. O propósito disso é permitir ao usuário que ele note, de uma maneira geral, o comportamento dos dados e, por meio de sua percepção especialista, identifique possíveis trechos interessantes para consulta ou análise.

Ao identificar as janelas de interesse e fornecer essa informação ao sistema, os dados são passados para o módulo descritor Coulomb juntamente com os dados brutos das séries. E após o processamento realizado pelo módulo descritor, o módulo responsável pela análise dos dados inicia seu trabalho.

O *data analysis module* (DAM) recebe como entrada os dados enviados pelo descritor Coulomb, os quais são compostos pelos intervalos pertencentes às séries que apresentam algum grau de similaridade com a janela de interesse passada pelo usuário. Desse ponto em diante, o DAM fica responsável por calcular a similaridade entre os intervalos existentes e a janela de interesse, utilizando para isso a função de distância Euclidiana. E com isso, ordena os elementos segundo o grau de similaridade obtido, passando ao VDEM os intervalos de maior relevância para serem exibidos para o usuário. Ou seja, o módulo DAM recebe como entrada os vetores de características e as subsequências juntamente com o nível de similaridade existente entre elas e retorna as subsequências mais similares que serão mostradas ao usuário pelo módulo VDEM.

Os experimentos visando testar o ambiente proposto, foram divididos em dois grupos distintos: i) experimentos com o descritor Coulomb e com o DAM para verificar o desempenho do descritor em reduzir a dimensionalidade dos dados e com o DAM para localizar as janelas da série com maior similaridade; ii) experimentos com o VDEM integrado aos demais módulos. Foram utilizadas bases geradas aleatoriamente, uma base de dados climáticos de diversas cidades brasileiras com temperatura mínima, máxima e índice de precipitação mensal dos anos de 1961 a 2010 obtidos em (AGRODATAMINE. . . , 2013) e dados médicos obtidos em (UCI. . . , 2013) com dados de nível de glicose de pacientes no decorrer de atividades diárias.

3.4.1 Experimentos e resultados obtidos

Para a realização dos experimentos do módulo de visualização e exploração dos dados foram feitas consultas por similaridade em séries temporais meteorológicas nos períodos de inverno ou verão. Ressalta-se ainda que esse tipo de base e consulta foi utilizada para que não se necessitasse de um especialista para verificar a eficácia do módulo em exibir intervalos similares.

A consulta exibida na figura 3.23 é uma consulta realizada na série temporal contendo a temperatura média da cidade de Araraquara/SP entre os anos de 1979 a 2010. Uma consulta *knn* com *n*= 10 e com o período de interesse correspondente ao inverno do ano de 1979 (período hachurado mais a esquerda do gráfico). E conforme nota-se na figura os períodos retornados (trechos hachurados do gráfico) pelo sistema correspondem a períodos de invernos de anos seguintes quando houve temperatura mínima próxima ao intervalo selecionado.

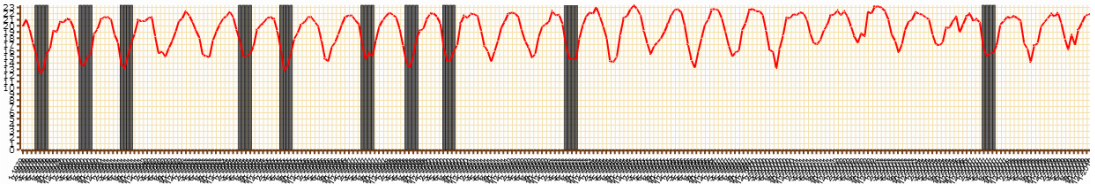

Figura 3.23: Consulta *knn* = 10 aos períodos de inverno da cidade de Araraquara/SP

Outro experimento utiliza três séries temporais referentes à temperatura máxima mensal das cidades de Avaré, São Paulo e Presidente Prudente dos anos de 1970 a 2008 e a consulta por similaridade, com *knn* = 10, é realizada selecionando, como período de interesse, o período de inverno da cidade de Presidente Prudente no ano de 1988. Conforme demonstrado na figura 3.24, os períodos de maior similaridade estão hachurados nos gráficos referentes às três séries temporais.

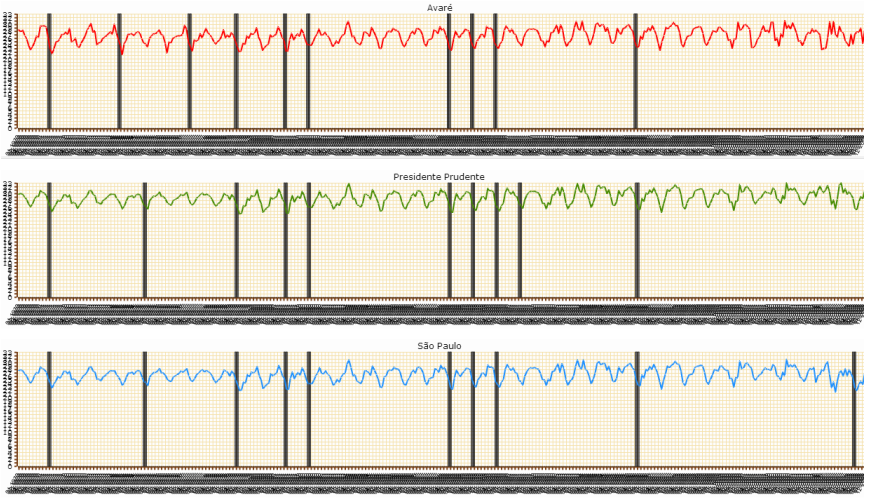

Figura 3.24: Consulta *knn* = 10 referente ao inverno de 1988 da cidade de Presidente Prudente/SP.

Conforme demonstrado, o módulo visual apresenta resultados satisfatórios, segundo nota-se pelos resultados das consultas visuais executadas e pela opinião do usuários que utilizaram o sistema, e permite ao especialista visualizar os intervalos similares de maneira inteligível e prática, proporcionando que as consultas por similaridade possam ser usadas para inferir conhecimento sobre as séries em análise.

3.5 Mineração de regras de associação

Com o intuito de conseguir minerar as séries temporais e obter regras de associação, foi proposto um sistema composto por módulos, o qual utiliza o descritor Coulomb em seu núcleo, que consegue gerar regras de associação para séries temporais, retornando uma resposta gráfica e textual ao usuário sobre os objetos de interesse pesquisados e a regras de associação geradas.

Seguem abaixo, as principais etapas para a mineração de regras de associação utilizando-se o descritor Coulomb. A Figura 3.25 mostra a interação entre os módulos e como o fluxo de dados ocorre dentro do método.

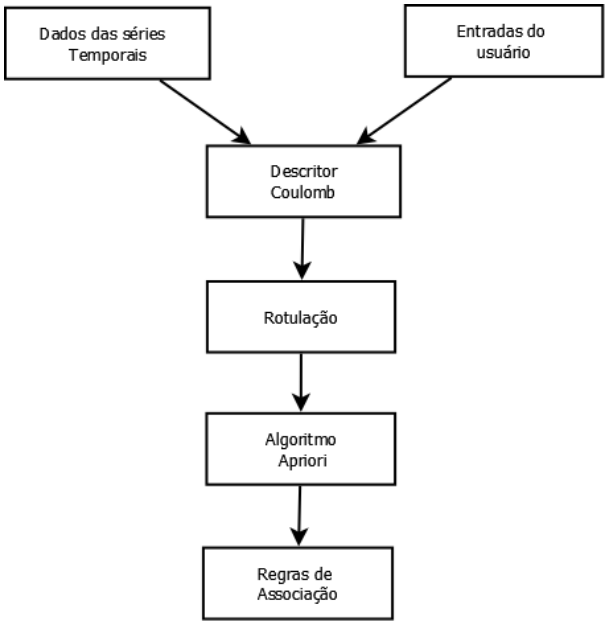

Figura 3.25: Esquema de interação dos módulos para a mineração de regras de associação.

Seleção e limpeza dos dados

Nessa etapa, as séries temporais são selecionadas e adequadas ao formato de entrada do método: $\{índice \longrightarrow observação\}$, em que o índice é o tempo em que foi obtida a observação. Para dados meteorológicos, por exemplo, ficaria: $\{mês \longrightarrow temperatura\}$ Além disso, os dados são verificados procurando-se dados inconsistentes ou incorretos. Outro fator importante é que as séries temporais podem conter falhas e essas precisam ser preenchidas para não causar distorções com relação a outras séries. Para a execução do método foi convencionado que em ocasiões de falhas da série essas falhas serão preenchidas com o valor $+\infty$, o que faz com que os dados não sejam utilizados para o cálculo da similaridade, ou descartados, caso a falha seja muito grande.

Redução do volume dos dados

Nessa etapa está inserido o descritor Coulomb para reduzir a dimensionalidade dos dados e encontrar intervalos similares. Nessa etapa, diferentemente de outros processos de mineração de dados, o especialista do domínio tem grande importância.

Após a escolha das séries temporais, também de acordo com o interesse do especialista, ele deve escolher as subsequências das séries temporais que são interessantes para a análise e, por consequência, para a geração das regras. Feito isso, o descritor Coulomb inicia a busca pelas subsequências com maior similaridade aos intervalos informados e até um limitante de similaridade informado também pelo especialista. Feito isso, os intervalos são rotulados para se tornarem compatíveis com o padrão de entrada do algoritmo de mineração. Além disso, os índices de tempo que compõem as subsequências são armazenados para posteriormente serem utilizados no cálculo da frequência de repetição. Essa frequência consegue verificar o índice de repetição mais frequente que a regra de associação ocorre.

Escolha do algoritmo de mineração de dados e extração de padrões

Para a geração de regras de associação foi utilizada uma versão do algoritmo *Apriori* (AGRAWAL; SRIKANT, 1994) em que os índices de suporte e confiança podem ser definidos pelo especialista ou, se não forem definidos, o algoritmo exibe todas as possibilidades de regras possíveis e seus índices.

Nessa fase são geradas as regras de associação, mas, no formato rotulado conforme feito pelo método de rotulação. Assim, a próxima etapa do método é traduzir os resultados obtidos pela mineração para que eles sejam apresentados ao usuário em formato de regras, e também, de visualizada num gráfico para que as regras fiquem mais inteligíveis para o usuário.

Nessa etapa também são realizados alguns cálculos estatísticos que oferecem ao especialista informações adicionais que o algoritmo *Apriori* não fornece em sua concepção original e que são importantes para a extração de conhecimento em séries temporais, como, por exemplo, a frequência de repetição de um dado padrão e a quantidade de ocorrências de determinado padrão na série.

Avaliação do conhecimento

Essa etapa é muito importante no processo de Mineração de Dados, na qual as regras extraídas são avaliadas e interpretadas e essa etapa é necessária para que o processo de descoberta

seja completo. Dessa maneira, é possível verificar a validade das regras obtidas, notadamente, verificando se houve a descoberta de conhecimento novo, útil e não trivial. Gerando, assim, conhecimento que pode ser utilizado para a tomada de decisões ou combinado com o conhecimento prévio do domínio e tornando mais compreensível ao usuário.

3.5.1 Experimentos e resultados obtidos

Para a realização dos experimentos foram utilizadas as seguintes séries temporais:

- Séries temporais da temperatura média mensal de sete aeroportos localizados em diferentes estados dos Estados Unidos do período de 1939 a 2011, obtidos em KMNI Climate Explorer (CLIMATE..., 2013). Nessa base foram selecionados, como interesse de consulta para a mineração, subsequências das séries temporais que apresentam picos de alta temperatura correspondentes ao inverno nos Estados Unidos.
- Séries temporais correspondentes à temperatura mínima, máxima e índice de precipitação mensal das cidades de Avaré, São Paulo e Presidente Prudente do estado de São Paulo, Brasil, obtidos da base de dados do projeto Agrodatamine (AGRODATAMINE..., 2013) dos anos de 1961 a 2010. Nessas séries foram selecionados, como subsequências de interesse, o períodos de alta pluviosidade e temperatura.

Como experimento inicial, foi realizada a mineração utilizando-se das séries temporais das temperaturas do Aeroporto Municipal de Bismarck, localizado no extremo norte dos Estados Unidos, e do Aeroporto de Nova York, localizado mais ao nordeste e com temperaturas maiores que o Aeroporto de Bismarck. No Aeroporto de Bismarck foi escolhido como subsequência de interesse o período de inverno norte-americano compreendido entre dezembro de um ano a março do próximo ano e no Aeroporto de Nova York, a subsequência relativa ao verão norte-americano, de junho a setembro de um ano.

A mineração trouxe as seguintes regras como resultado:

Regras geradas:

- S0P1 ← S0P0 (32.2222, 100). Repetição: 11
- S0P0 ← S0P1 (32.2222, 100). Repetição: 11
- S1P1 ← S1P0 (17.7778, 100). Repetição: 12
- S1P0 ← S1P1 (17.7778, 100). Repetição: 12

Figura 3.26: Regras geradas para a base de Aeroportos.

Em que $S0$ corresponde à série do Aeroporto de Bismarck e $S1$ corresponde à série do Aeroporto de Nova York. O padrão $P0$ corresponde à subsequência referente ao inverno em Bismarck no ano de 1939/1940 e o padrão $P1$ corresponde à subsequência referente ao verão em Nova York no ano de 1940.

Como exemplo, a interpretação da regra: $S0P1 \leftarrow S0P0(32.2222, 100)$. *Repetição: 11*, significa que quando ocorre o inverno em Bismark ocorre também o verão em Bismark com uma confiança de 100% e com um suporte de 32,22% e, pelas medidas de frequência analisadas, esse padrão ocorre a cada 11 meses. As outras regras podem ser interpretadas no mesmo sentido. Assim, verifica-se a eficácia do método para encontrar a regra de associação validadas e com a informação adicional da repetição dada pelo próprio método e validada por um especialista em meteorologia, afirmando que a variação de 1 mês entre as estações climáticas é aceitável para séries com temperatura mensal.

Outro experimento realizado utilizou as séries temporais de temperatura máxima da cidade de São Paulo e a precipitação pluviométrica da cidade. As subsequências de interesse para análise foram: $P0$ período com maior pluviosidade correspondente aos meses de dezembro/1962 a março/1963; $P1$ período correspondente ao verão brasileiro correspondente aos meses de dezembro/1962 a março/1963; e $P2$ período correspondente ao inverno brasileiro correspondente aos meses de junho/1962 a setembro/1962; a série $S0$ corresponde à série temporal da precipitação; e a $S1$ corresponde à série com a temperatura máxima. As regras de maior interesse obtidas foram:

Regras geradas:

- S0P1 ← S0P0 (33.0645, 100). Repetição: 13
- S0P0 ← S0P1 (33.0645, 100). Repetição: 13
- S1P0 ← S1P1 S1P2 (12.9032, 100). Repetição: 48
- S1P2 ← S1P1 (12.9032, 100). Repetição: 48
- S1P1 ← S1P2 (12.9032, 100). Repetição: 48
- S0P2 ← S0P0 S0P1 (33.0645, 100). Repetição: 13
- S0P1 ← S0P0 S0P2 (33.0645, 100). Repetição: 13
- S0P0 ← S0P1 S0P2 (33.0645, 100). Repetição: 13

Figura 3.27: Regras geradas para a base Agrodatamine.

Assim, depreende-se da primeira regra, como exemplo, que quando chove ocorre também o verão com confiança de 100% e suporte de 33,06% e essa condição se repete a cada 13 meses. Dessa mesma maneira as outras regras podem ser interpretadas.

Um fato interessante são as regras que apresentam repetição a cada 48 meses. Essas regras, segundo o especialista, ocorreram devido ao fato que as características climáticas do ano de

1962 apresentarem características peculiares que ocorrem devido à influência do fenômeno *El niño* em que as características climáticas se repetem a cada 12 anos.

Dessa forma, é possível validar o método de mineração de regras de associação para descoberta de conhecimento não trivial existente nas séries temporais.

3.6 Considerações Finais

Através da análise dos resultados obtidos, conclui-se que o descritor Coulomb apresenta acurácia e tempo satisfatórios para a execução de consultas por similaridade em séries temporais uni e multidimensionais. Além disso, na comparação do método Coulomb com os métodos tradicionais de busca em séries temporais, por meio da análise dos gráficos de precisão x revocação, nota-se um expressivo ganho. Isso faz do método Coulomb um potencial descritor para análise de séries temporais e,conforme visto, viabiliza a execução de consultas visuais e a mineração em séries. Além disso, o método para mineração de regras de associação em séries temporais apresenta resultados satisfatórios com relação às regras geradas.

Capítulo 4

CONCLUSÕES

Este capítulo apresenta o trabalho desenvolvido até o presente momento. O capítulo está organizado da seguinte maneira: a Seção 4.1 apresenta as considerações iniciais sobre o trabalho desenvolvido; na Seção 4.2 é apresentada a relação de atividades desenvolvidas e os resultados preliminares obtidos; e a Seção 4.3 apresenta as considerações finais do capítulo.

4.1 Considerações Iniciais

De acordo com a proposta elaborada na seção anterior, preliminarmente, foram estudados os principiais conceitos relacionados à execução do projeto, como: mineração em dados complexos, descritores de séries, métodos de validação da eficiência de descritores, entre outros conceitos e, também, foram realizadas atividades visando analisar a viabilidade de execução do projeto como um todo. E isso culminou com a execução da primeira etapa do projeto com a proposta de um novo descritor para série temporal. Nas seções seguintes serão apresentadas breves descrições dos trabalhos iniciais juntamente com o descritor desenvolvido e os experimentos utilizados para a sua validação.

4.2 Contribuições

Neste trabalho, o principal objetivo foi a elaboração de um método que pudesse ser utilizado para realizar consultas por similaridade em séries temporais uni e multidimensionais e que diminuísse a complexidade computacional e aumentasse a acurácia. Assim, a principal contribuição deste trabalho foi a elaboração de um método que atendesse tais requisitos e que pudesse contribuir para a realização de consultas por similaridade em séries temporais. Além

disso, diversas outras contribuições foram obtidas durante o progresso do trabalho e colaboraram para a agregação de valor ao objetivo inicial.

Juntamente com o descritor proposto, foi desenvolvido, também, um sistema para a execução de consultas visuais, permitindo a interação com o usuário, e apresentando as respostas graficamente. Isso contribuiu para validar o método proposto e tornar mais simples e intuitivo a resposta dada pelo sistema ao usuário. Outra contribuição foi o acoplamento do descritor Coulomb para ser utilizado no processo de mineração de regras de associação.

Além disso, o estudo realizado sobre o tema ”Consultas por similaridade em séries temporais”, fundamental para a realização de todas as atividades envolvidas, foi importante para verificar a importância do tema e da necessidade de cobrir uma lacuna existente na área de encontrar um descritor que apresente um boa acurácia e baixa complexidade computacional.

A partir da realização do levantamento por meio do uso de técnicas de revisão sistemática, pôde-se concluir que a consulta por similaridade em séries temporais tem sido objeto de estudo de muitas pesquisas, e que elas têm crescido em número e abrangência nos últimos anos. Contudo, este levantamento também mostrou que, do ponto de vista de acurácia e complexidade computacional, ainda há muito o que ser explorado para desenvolver técnicas suficientemente satisfatórias nesses quesitos.

Finalmente, a partir dos resultados alcançados por meio dos estudos e implementações realizados, foi possível obter conclusões interessantes sobre a consulta por similaridade em séries temporais. A análise dos resultados obtidos confirmam diversas premissas, principalmente de que a manipulação de séries temporais é um processo custoso e que exige a utilização de descritores para prover uma resposta eficiente e com certa acurácia.

A realização deste trabalho de mestrado também foi muito importante para contribuir com a formação do autor como pesquisador. Adquirir conhecimento técnico para uso do ferramental científico e o aprendizado com a busca por soluções para os problemas encontrados colaboraram muito para o crescimento de habilidades para o desenvolvimento de pesquisas futuras em nível de doutorado.

4.3 Trabalhos futuros

A realização deste trabalho criou diversas perspectivas para a elaboração de trabalhos futuros como continuação desta pesquisa. Por meio da revisão sistemática realizada, notou-se que um campo de pesquisa extremamente vasto e inexplorado no que tange a séries temporais são

as séries espaço-temporais. Como trabalho futuro, fica a proposta de integração do descritor proposto para a realização de consultas por similaridade em séries espaço-temporais. Notou-se também que não existe na literatura uma métrica satisfatória e suficiente para medir a eficácia de um descritor. Logo a proposta de uma métrica desse tipo pode ser explorada como trabalho futuro.

4.4 Produção científica

4.4.1 Artigos em periódicos e anais de eventos

- Andrade, C.G.; Ribeiro, M.X.; Yaguinuma, C.A.; Santos, M.T.P. A Novel Method for Similarity Search over Meteorological Time Series Data based on the Coulomb's Law. In: *ICEIS 2013 - Proceedings of the 15th International Conference on Enterprise Information Systems*. Volume 1, Angers, France, SciTePress, 4-7 July, 2013;

- Andrade, C.G.; Ribeiro, M.X. Searching for similarities in series using Coulomb's law. In: *KDMiLe'13 - Symposium on Knowledge Discovery, Mining and Learning*. São Paulo, Brazil, 17-19 July, 2013;

- Andrade, C.G.; Ribeiro, M.X. Similarity Search in multidimensional time series using the Coulomb's law. In: *Journal of Information And Data Management*. Special Issue, pags. 74-83, Brazil, 2014;

- Andrade, C.G.; Ribeiro, M.X. A similarity searching-based method for visual search in time series using Coulomb's law. In: *SISAP 2014 - 7th International Conference on Similarity Search and Applications*, pags. 241-246, Los Cabos, Mexico, 29-31 October, 2014;

- Andrade, C.G.; Cazzolato, M. T.; Ribeiro, M.X. Data Mining in Meteorological Time Series using Association Rules and a Similarity Searching-Based Method. In: *2nd KDMiLe - II Symposium on Knowledge Discovery, Mining and Learning*. São Paulo, Brazil, 20-21 October, 2014;

4.4.2 Outras publicações geradas durante o mestrado

- Andrade, C.G.; Kawakami, C.; Betetto, L.A.O.; Ribeiro, M.X. A proposal for measuring interest in privacy preservation in data mining using Jaccard index. In: *KDMiLe'13* -

Symposium on Knowledge Discovery, Mining and Learning. São Paulo, Brazil, 17-19 July, 2013.

4.5 Considerações Finais

O projeto de mestrado elaborou uma solução para o problema descrito e caracterizado na proposta de qualificação e, como fruto da resolução desse problema, foi elaborado um descritor para consultas por similaridade em séries temporais. Os resultados do andamento da pesquisa foram analisados e avaliados pela comunidade científica da área e culminaram na publicação de artigos científicos em conferências e periódicos. Demonstrando assim, que o objetivo primordial do mestrado, que é apresentar uma solução para um problema utilizando-se do ferramental disponível, foi atingido com êxito.

REFERÊNCIAS BIBLIOGRÁFICAS

AGRAWAL, R.; FALOUTSOS, C.; SWAMI, A. N. Efficient similarity search in sequence databases. In: *Proceedings of the 4th International Conference on Foundations of Data Organization and Algorithms*. London, UK, UK: Springer-Verlag, 1993. (FODO '93), p. 69–84. ISBN 3-540-57301-1. Disponível em: <http://dl.acm.org/citation.cfm?id=645415.652239>.

AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining association rules in large databases. In: *Proceedings of the 20th International Conference on Very Large Data Bases*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994. (VLDB '94), p. 487–499. ISBN 1-55860-153-8. Disponível em: <http://dl.acm.org/citation.cfm?id=645920.672836>.

AGRODATAMINE: Development of Algorithms and Methods of Data Mining to Support Researches on Climate Changes Regarding Agrometeorology. [S.l.], 2013. `http://www.gbdi.icmc.usp.br/projects/agrodatamine/index.html`.

BARIONI, M. C. N. *Operações de consulta por similaridade em grandes bases de dados complexos*. Tese (Doutorado) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2006.

BERNDT, D. J.; CLIFFORD, J. Using dynamic time warping to find patterns in time series. In: *KDD Workshop*. [S.l.: s.n.], 1994. p. 359–370.

BOZKAYA, T.; OZSOYOGLU, M. Indexing large metric spaces for similarity search queries. *ACM Trans. Database Syst.*, ACM, New York, NY, USA, v. 24, n. 3, p. 361–404, set. 1999. ISSN 0362-5915. Disponível em: <http://doi.acm.org/10.1145/328939.328959>.

BUSSAB, W. de O.; MORETTIN, P. *Estatística básica*. [S.l.]: Saraiva, 2008. ISBN 9788502034976.

CAMERRA, A. et al. isax 2.0: Indexing and mining one billion time series. In: *Proceedings of the 2010 IEEE International Conference on Data Mining*. Washington, DC, USA: IEEE Computer Society, 2010. (ICDM '10), p. 58–67. ISBN 978-0-7695-4256-0. Disponível em: <http://dx.doi.org/10.1109/ICDM.2010.124>.

CHAKRABARTI, K. et al. Locally adaptive dimensionality reduction for indexing large time series databases. *ACM Trans. Database Syst.*, ACM, New York, NY, USA, v. 27, n. 2, p. 188–228, jun. 2002. ISSN 0362-5915. Disponível em: <http://doi.acm.org/10.1145/568518.568520>.

CHAN, K.-P.; FU, A.-C. Efficient time series matching by wavelets. In: *Data Engineering, 1999. Proceedings., 15th International Conference on*. [S.l.: s.n.], 1999. p. 126–133. ISSN 1063-6382.

CLIMATE-PREDICTION-CENTER. 2012. Disponível em: <http://www.cpc.ncep.noaa.gov/products/analysis-monitoring/ensostuff/ONI-change.shtml>.

CLIMATE Prediction Center. 2013.

ELMASRI, R.; NAVATHE, S. *Sistemas de banco de dados*. Pearson Addison Wesley, 2006. ISBN 9788588639171. Disponível em: <http://books.google.com.br/books?id=tylQGgAACAAJ>.

FALOUTSOS, C.; RANGANATHAN, M.; MANOLOPOULOS, Y. Fast subsequence matching in time-series databases. In: *Proceedings of the 1994 ACM SIGMOD international conference on Management of data*. New York, NY, USA: ACM, 1994. (SIGMOD '94), p. 419–429. ISBN 0-89791-639-5. Disponível em: <http://doi.acm.org/10.1145/191839.191925>.

FAYYAD, U. et al. Knowledge discovery and data mining: Towards a unifying framework. In: . [S.l.]: AAAI Press, 1996. p. 82–88.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI Magazine*, v. 17, p. 37–54, 1996.

FERREIRA, M. R. P. et al. Adding knowledge extracted by association rules into similarity queries. In: . [S.l.]: Journal of Information and Data Management, 2010. p. 391–406.

KEOGH, E. A fast and robust method for pattern matching in time series databases. In: *Proceedings of WUSS-97*. [S.l.: s.n.], 1997.

KEOGH, E. et al. Dimensionality reduction for fast similarity search in large time series databases. *Knowledge and Information Systems*, Springer-Verlag London Limited, v. 3, n. 3, p. 263–286, 2001. ISSN 0219-1377. Disponível em: <http://dx.doi.org/10.1007/PL00011669>.

KMNI Climate Explorer. 2013.

KORN, F.; JAGADISH, H. V.; FALOUTSOS, C. Efficiently supporting ad hoc queries in large datasets of time sequences. *SIGMOD Rec.*, ACM, New York, NY, USA, v. 26, n. 2, p. 289–300, jun. 1997. ISSN 0163-5808. Disponível em: <http://doi.acm.org/10.1145/253262.253332>.

LIN, J. et al. A symbolic representation of time series, with implications for streaming algorithms. In: *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*. New York, NY, USA: ACM, 2003. (DMKD '03), p. 2–11. Disponível em: <http://doi.acm.org/10.1145/882082.882086>.

MEADOW, C.; BOYCE, B.; KRAFT, D. *Text information retrieval systems*. Academic Press, 2000. (Library and Information Science Series). ISBN 9780124874053. Disponível em: <http://books.google.com.br/books?id=pWgVAQAAIAAJ>.

MORINAKA, Y. et al. *The L-index: An Indexing Structure for Efficient Subsequence Matching in TimeSequence Databases*. 2001. 51-60 p.

PARIS, A. royale des sciences. *Histoire de l'Academie royale des sciences*. De l'imprimerie royale, 1788. Disponível em: <http://books.google.com.br/books?id=by5EAAAAcAAJ>.

PENATTI, O. A. B. *Estudo comparativo de descritores para recuperação de imagens por conteudo na web*. Tese (Doutorado) — Universidade Estadual de Campinas, Instituto de Computação, 2009.

RIBEIRO, M. X. *Suporte a Sistemas de Auxílio ao Diagnóstico e de Recuperação de Imagens por Conteúdo Usando Mineração de Regras de Associação*. Tese (Doutorado) — Instituto de Ciências Matemáticas e de Computação - ICMC, USP, São Carlos, 2008.

SANTOS, I. J. P. d. *TRACTS : um método para classificação de trajetórias de objetos móveis usando séries temporais*. Tese (Doutorado) — Universidade Federal do Rio Grande do Sul. Instituto de Informática, 2011.

SRIKANT, R.; AGRAWAL, R. Mining quantitative association rules in large relational tables. *SIGMOD Rec.*, ACM, New York, NY, USA, v. 25, n. 2, p. 1–12, jun. 1996. ISSN 0163-5808. Disponível em: <http://doi.acm.org/10.1145/235968.233311>.

STANDARDIZATION, I. O. for; 69, T. C. I. *Accuracy (trueness and Precision) of Measurement Methods and Results: Exactitude (justesse Et Fidélité) Des Résultats Et Méthodes de Mesure. Partie 2, Méthode de Base Puor la Détermination de la Répétabilité Et de la Reproductibilité D'une Méthode de Mesure Normalisée. Basic method for the determination of repeatability and reproducibility of a standard measurement method. Part 2*. International Organization for Standardization, 1994. (International standard). Disponível em: <http://books.google.com.br/books?id=nSJnPAAACAAJ>.

TANAKA, Y.; IWAMOTO, K.; UEHARA, K. Discovery of time-series motif from multi-dimensional data based on mdl principle. *Mach. Learn.*, Kluwer Academic Publishers, Hingham, MA, USA, v. 58, n. 2-3, p. 269–300, fev. 2005. ISSN 0885-6125. Disponível em: <http://dx.doi.org/10.1007/s10994-005-5829-2>.

TORRES, R. D. S.; FALCÃO, A. X. Content-based image retrieval: Theory and applications. *Revista de Informática Teórica e Aplicada*, v. 13, p. 161–185, 2006.

TOSCANI, L.; VELOSO, P. *COMPLEXIDADE DE ALGORITMOS*. [S.l.]: BOOKMAN COMPANHIA ED, 2008. ISBN 9788577803507.

UCI Machine Learning Repository: Diabetes Data Set. 2013.

WEI, W. *Time series analysis: univariate and multivariate methods*. [S.l.]: Pearson Addison Wesley, 2006. ISBN 9780321322166.