# Avaliação do Algoritmo de Força-Bruta para a Identificação de Padrões Freqüentes em Séries Temporais*

**Daniel Moreira Cestari**[1], **André Gustavo Maletzke**[1,2,3], **Gustavo E.A.P.A Batista**[1,3]

[1]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo
Laboratório de Inteligência Computacional – LABIC
Caixa Postal 668, 13560-970 – São Carlos, SP, Brasil

[2]Centro de Engenharias e Ciências Exatas – Universidade Estadual do Oeste do Paraná
Laboratório de Bioinformática – LABI
Parque Tecnológico Itaipu – PTI
Caixa Postal 39, 85856-970 – Foz do Iguaçu, PR, Brasil

[3]Centro de Estudos Avançados em Segurança de Barragens – CEASB
Parque Tecnológico Itaipu – PTI

cestari@grad.icmc.usp.br, {andregm,gbatista}@icmc.usp.br

***Abstract.*** *There exists an increasing interest in discovering knowledge from temporal data. This interest is supported by the large number of practical problems that can only be completely analyzed when the time dimension is considered. An approach used to analyze temporal data is to discover frequent patterns that can be used to describe a time series. This work provides a study regarding the brute-force approach to identify frequent patterns in time series data. Our studies indicate that this approach has quadratic complexity and is limited to be applied to small-sized time series.*

***Resumo.*** *Existe atualmente um interesse crescente por descobrir conhecimento a partir de dados que possuem uma característica temporal. Esse interesse é proveniente do grande número problemas práticos que somente podem ser analisados por completo quando a dimensão tempo é considerada. Uma abordagem utilizada para analisar dados temporais é a descoberta de padrões freqüentes que podem ser utilizados para descrever a série temporal. Neste trabalho é realizado um estudo sobre a abordagem força-bruta para a identificação de padrões freqüentes em uma série temporal. Os estudos indicam que essa abordagem possui complexidade quadrática de tempo de execução, e que essa abordagem somente pode ser indicada para séries temporiais de tamanho reduzido.*

## 1. Introdução

O surgimento de novas tecnologias tem possibilitado acumular informações nas mais variadas áreas, dificultando a análise desse grande volume de dados através dos métodos e ferramentas tradicionais. Desse modo, cada vez mais métodos, processos e ferramentas computacionais são utilizadas para auxiliar a especialistas de distintos domínios na análise desses grandes conjuntos de informações. No entanto, para que essas técnicas

*Trabalho desenvolvido com o apoio do Programa de Desenvolvimento Tecnológico Avançado – PDTA-FPTI/BR

computacionais possam ser aplicadas é necessário que esses dados estejam dispostos em um formato estruturado adequado [Rezende 2003].

Atualmente, registram-se dados nos mais variados formatos e tipos. Dentre os quais, grande parte provenientes de processos de monitoração realizados ao longo do tempo, podendo caracterizar um tipo de dado denominado de Série Temporal – ST. Uma ST pode ser definida como sendo uma coleção de observações, de um determinado fenômeno, realizadas ao longo do tempo [Morettin and Toloi 2006].

Diversos processos computacionais que auxiliam na análise de grandes conjuntos de dados possuem aplicação restrita à dados que não possuem uma relação temporal entre os fatos existentes. Entretanto, existem aplicações em que a característica temporal e/ou seqüencial existente nos dados é importante para a análise, pois pode conter informações cruciais para a identificação e compreensão de possíveis fenômenos existentes nos dados [Fink 2004]. Esse fato tem impulsionado a comunidade acadêmica no estudo e desenvolvimento de métodos que permitam a exploração desses dados.

Uma das abordagens que tem recebido cada vez mais atenção refere-se à identificação de padrões existentes em uma ST, de modo que esses padrões possam descrever essa série. A identificação de sub-seqüências repetidas dentro de uma ST pode ser utilizada como um padrão para descrever essa ST [Tatavarty et al. 2007]. Esses padrões são denominados de padrões freqüentes e na área de estudo do seqüenciamento genético são conhecidos como *motifs*, por se tratarem de cadeias compostas por um alfabeto determinado e limitado. Na Figura 1 é apresentada uma série temporal na qual foi identificado três ocorrências de um determinado padrão nas posições A,B e C.

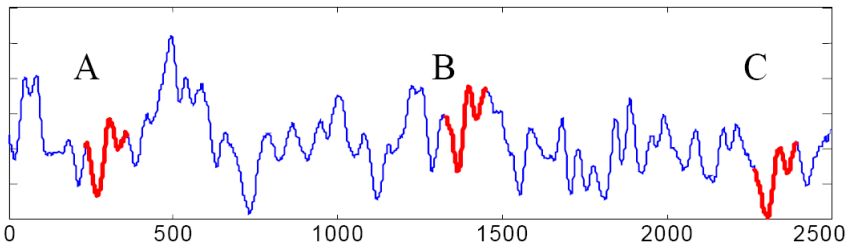


**Figura 1. Exemplo de um padrão freqüente identificado em uma série temporal [Lin et al. 2002].**

Desse modo, neste trabalho é apresentado um estudo preliminar de um método para identificação de padrões freqüentes em ST denominado de força-bruta. É apresentada, posteriormente, uma avaliação do método estudado por meio de ST geradas artificialmente.

Este artigo está organizado da seguinte maneira: na Seção 2 é descrito o algoritmo de identificação de padrões freqüentes, bem como algumas heurísticas simples utilizadas para melhorar o seu desempenho; na Seção 3 são apresentados alguns resultados experimentais preliminares; por fim, na Seção 4 são apresentadas as conclusões e possibilidades de trabalhos futuros.

## 2. Materiais e Métodos

A identificação de padrões freqüentes dentro de uma série temporal tem sido cada vez mais utilizada na mineração de ST de distintos domínios [Chiu et al. 2003]. Diversas abordagens podem ser aplicadas para a identificação desses padrões. Neste trabalho é estudado e avaliado o método de identificação denominado de força-bruta (*brute-force*).

Para melhor entendimento dos conceitos relacionados ao método de identificação de padrões freqüentes são apresentadas algumas definições de termos utilizados neste trabalho.

**Definição 1** *Série Temporal [Chiu et al. 2003]: considera-se uma série temporal $T$ um conjunto ordenado de tamanho $N$, sendo $T = (x_1, \ldots, x_N)$, na qual $N$ é o número total de observações.*

É importante observar que uma ST pode conter uma quantidade grande de observações, e, geralmente não existe o interesse de se minerar características globais da série, mas sim a identificação de pequenas porções existentes dentro da série denominadas de sub-seqüências.

**Definição 2** *Subseqüência [Chiu et al. 2003]: uma sub-seqüência $C$ de uma ST $T$ é uma amostra contínua de $T$ de tamanho $n$, sendo $n << N$, isto é, $C = (x_p, ..., x_{p+n-1})$ para $1 \leq p \leq N - n + 1$.*

Outro conceito, amplamente, utilizado na identificação de padrões freqüentes refere-se à extração de todas as sub-seqüências existentes na ST, a qual pode ser realizada por meio de uma janela deslizante, cujo tamanho deverá ser igual ao tamanho das sub-seqüências que deseja-se extrair.

**Definição 3** *Janela Deslizante: Dada a série temporal $T$ de tamanho $N$ e um tamanho de sub-seqüência $n$ previamente determinado pelo usuário ou especialista do domínio, é construída uma matriz $M$ com todas as possíveis sub-seqüências. Essas sub-seqüências podem ser construídas por meio do deslocamento, sobre $T$, de uma janela de tamanho $n$ inserindo na posição $p$ da matriz a sub-seqüência $C_p$, sendo que o tamanho da matriz $M$ é $(N - n + 1)$ por $n$.*

A identificação de padrões freqüentes por meio do algoritmo de força-bruta é uma das abordagens mais simples utilizadas na mineração de ST. Essa abordagem possui alto custo computacional e possui uma relação quadrática com o tamanho da série que será minerada [Lin et al. 2002]. No entanto, apresenta resultados satisfatórios para ST de baixa dimensionalidade. No Algoritmo 1 é apresentado o pseudo-código da abordagem por força-bruta.

Inicialmente, é necessário definir o tamanho dos padrões que se deseja buscar na série temporal a ser minerada. No Algoritmo 1 busca-se identificar um padrão de tamanho $n$ em uma série temporal de tamanho $N$. A variável $i$ controla o início da sub-seqüência $S_i$ de tamanho $n$ que se deseja verificar se é um padrão freqüente. Por meio da variável $i$ é implementada a janela deslizante que irá percorer toda a ST. A variável $j$ controla o início sub-seqüência $S_j$ que será comparada com $S_i$. Se a distância, $D$, entre $S_i$ e $S_j$ for menor que um limiar $R$ então $S_i$ e $S_j$ são consideradas similares e as posições de casamento $i$ e $j$ são armazenadas em um conjunto $casamentos$.

**Algoritmo 1** Algoritmo força-bruta para identificação de padrões freqüentes.

```
1: for i = 1 to N − n + 1 do
2:     casamentos ⇐ ∅
3:     for j = i + 1 to N − n + 1 do
4:         if (D(S_i, S_j) < R) then
5:             casamentos ⇐ casamentos ∪ (i, j)
6:         end if
7:     end for
8: end for
```

Neste trabalho, para realizar a comparação entre duas sub-seqüências é utilizada a distância Euclidiana, a qual é descrita na Equação 1. Na Equação 1 é realizada a comparação entre as sub-seqüências $S_i$ e $S_j$, ambas de tamanho $n$.

$$D(S_i, S_j) = \left( \sum_{k=1}^{n} (S_{ik} - S_{jk})^2 \right)^{\frac{1}{2}} \quad (1)$$

na qual $S_{ik}$ e $S_{jk}$ são as $k$-ésimas observações das sub-seqüências $S_i$ e $S_k$, respectivamente.

O conceito de casamento (*match*) é simples e intuitivo, no entanto, para que seja possível a identificação de padrões freqüentes em uma série temporal é necessário identificar e descartar situações nas quais ocorrem casamentos triviais (*trivial matches*).

**Definição 4** *Casamento trivial [Chiu et al. 2003]: dada uma série temporal contendo as sub-seqüências $S_i$ e $S_j$ começando em $i$ e $j$, respectivamente, e ocorra um casamento entre $S_i$ e $S_j$, considera-se um casamento trivial de $S_i$ com $S_j$ se $i = j$ ou se não existe uma sub-seqüência $S_k$ começando em $k$ tal que $D(S_i, S_k) > R$, para $i < k < j$ ou para $j < k < i$.*

Casamentos triviais são um problema importante na identificação de padrões freqüentes. Isso porque quando existe um casamento entre duas sub-seqüências em uma posição $p$, normalmente existem alguns casamentos triviais com sub-seqüências localizadas em posições próximas a $p$, tanto à direita quanto à esquerda de $p$.

Como mencionado, a abordagem por força-bruta requer uma complexidade de tempo elevada devido à quantidade de comparações que são realizadas durante todo o processo, sendo a complexidade de tempo do algoritmo $O(N^2 n)$, onde $N$ é o tamanho da ST e $n$ é o tamanho da sub-seqüência procurada.

Algumas heurísticas podem ser utilizadas com o intuito de reduzir os fatores constantes da complexidade de tempo de execução dessa abordagem. Neste trabalho foram utilizadas quatro heurísticas:

- **Heurística 1 — comparar somente com posições maiores que o início da sub-seqüência:** uma sub-seqüência que inicia na posição $i$ é comparada somente com posições da ST maiores do que $i$ (veja Algoritmo 1, linha 3). O fato de não comparar com sub-seqüências que iniciam em posições menores do que $i$, além de eliminar casamentos triviais, reduz a complexidade computacional;

- **Heurística 2 — selecionar a menor distância:** As sub-seqüências próximas ao padrão tendem a ter uma distância menor, conseqüentemente são interpretadas também como um padrão freqüente, embora sejam casamentos triviais. Para eliminar essa redundância, no cálculo da distância entre sub-seqüências são consideradas somente as sub-seqüências cuja distância é um mínimo local. Assim, durante a execução, enquanto o valor da distância diminui não é considerada essa sub-seqüência como padrão. Somente quando a distância começar a crescer será considerada como tal;

Quando a janela deslizante está próxima a um padrão, a distância entre o padrão e as sub-seqüências adjacentes é menor que o limiar especificado, logo elas são consideradas como padrão. Mas não há interesse nesses casamentos triviais que ocorrem nas proximidades do padrão. Esta heurística elimina essa redundância, considerando como padrão apenas a sub-seqüência com a menor distância, segundo explicado acima.

- **Heurística 3 — interromper o cálculo da distância:** uma vez que a medida calculada é a distância euclidiana, os valores intermediários do cálculo da distância são comparados com os valores do limiar. Caso o valor intermediário ultrapasse o valor do limiar, o cálculo da distância é interrompido, já que não pode haver casamento;
- **Heurística 4 — *offset translation*:** consiste, basicamente, em substituir cada amostra da sub-seqüência pela subtração da respectiva amostra pela média da sub-seqüência, de modo que a sub-seqüência permaneça na origem em relação ao eixo das ordenadas. Evitando assim que um padrão deslocado em relação ao eixo das ordenadas não seja identificado.

Na Figura 2 é apresentado um exemplo de utilizacao da heurística 4, na qual a esquerda vemos duas sub-seqüências originais da ST, e a direita vemos as duas sub-seqüências após a aplicação do *offset translation*. O cálculo da distância entre as sub-seqüências é feito após a aplicação do *offset translation*. Por meio da Figura 2 vemos a importância dessa heurística, pois se o cálculo da distância fosse realizado sem a aplicação da heurística, padrões neste formato não seriam encontrados.

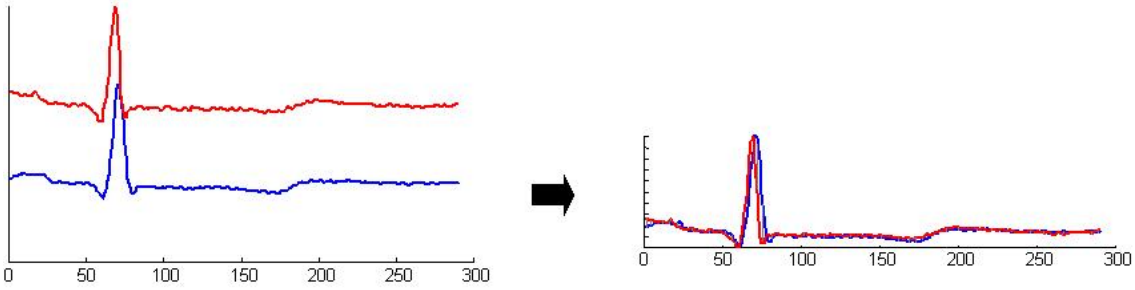


**Figura 2. Exemplo de utilização do *offset translation***

## 3. Avaliação

Com objetivo da avaliar a abordagem por força-bruta foram utilizadas as séries temporais *Random Walk*, *Tide* e *Burst* [Lin et al. 2002]. Na preparação dos experimentos foram realizados os passos descritos a seguir:

- Para cada série temporal foi extraída uma sub-seqüência contendo as primeiras observações da série. Essa sub-seqüência, denominada $T$, possui 500 observações para a série *Random Walk*, 1000 observações para a ST *Tide* e 1500 observações para a ST *Burst*. A sub-seqüência $T$ é utilizada para receber o padrão inserido artificialmente;
- Para cada série temporal foi extraída uma sub-seqüência $P$ que será utilizada como padrão. Para a série *Random Walk*, $P$ possui 50 observações, para a ST *Tide* $P$ possui 100 observações e para a ST *Burst* $P$ possui 150 observações. Para cada série existe a restrição de que $P$ não esteja compreendida entre as primeiras posições que constituem a sub-série $T$;
- Foram inseridas duas ocorrências da sub-seqüência $P$ na sub-série $T$. A inserção de ambas as ocorrências foi realizada em posições aleatórias, somente restringindo sobreposição entre cada sub-seqüência inserida. É importante ressaltar que o processo de inserção não substitui nenhuma amostra da sub-série $T$. A inserção das sub-seqüências foi realizada de modo a acompanhar o comportamento da sub-série $T$, evitando diferenças abruptas entre o ponto de inserção da sub-série $T$ e primeiro ponto da sub-seqüência $P$.

Esse processo foi repetido dez vezes considerando posições distintas de inserção das sub-seqüências para cada rodada. Assim, obteve-se, para cada ST, um conjunto de dez séries $T$ com duas ocorrências da sub-seqüência $P$ inseridas em posições distintas.

Foi executado o algoritmo de força-bruta com as heurísticas mencionadas anteriormente. Esse algoritmo requer que alguns parâmetros sejam definidos. Neste experimento foram utilizados os seguintes parâmetros: tamanho do padrão a ser identificado igual a 10% da ST $T$ antes da inserção das duas sub-seqüências e limiar de similaridade de $0.08$, considerando que antes da comparação das duas sub-seqüências, ambas sejam normalizadas para média igual a 0 e desvio-padrão igual a 1.

Os resultados indicam que todos os padrões foram encontrados pelo algoritmo força-bruta nas 10 execuções realizadas para cada ST. Esse resultado era esperado pois o algoritmo força-bruta realiza todas as comparações entre o padrão e as sub-seqüências extraídas pela janela deslizante. Entretanto, como mencionado anteriormente, esse grande número de comparações é bastante custoso. A Tabela 1 mostra os tempos de execução médio para cada ST.

**Tabela 1. Tempo médio de execução do algoritmo de força-bruta**

| Série Temporal | Tempo Médio de Execução (min.) | Desvio Padrão |
|---|---|---|
| *Random Walk* | 1.437 | 0.003 |
| *Tide* | 10.055 | 0.004 |
| *Burst* | 33.004 | 0.007 |

Os tempos apresentados na Tabela 1 comprovam experimentalmente que o algoritmo possui complexidade quadrática. Essa complexidade permite que o algoritmo seja aplicado somente à pequenas séries temporais. A ST *Random Walk*, com apenas 500 observações, teve um tempo de execução aceitável. Entretanto, a série *Burst* com 1500 observações levou mais de 30 minutos para cada repetição do experimento.

Embora o algoritmo de força-bruta seja muito ineficiente para o uso em problemas práticos que tenham grandes entradas, a implementação desse algoritmo possui grande aplicabilidade científica. O objetivo de implementar esse algoritmo é de utilizá-lo como base para comparação com outros algoritmos mais eficientes, mas que podem não identificar todos os padrões presentes na ST.

Por fim, na Figura 3 são apresentadas graficamente as primeiras 500 observações da ST *Random Walk* juntamente com os dois padrões inseridos aleatoriamente e identificados corretamente por meio do processo de identificação de padrões freqüentes realizado pela abordagem por força-bruta.

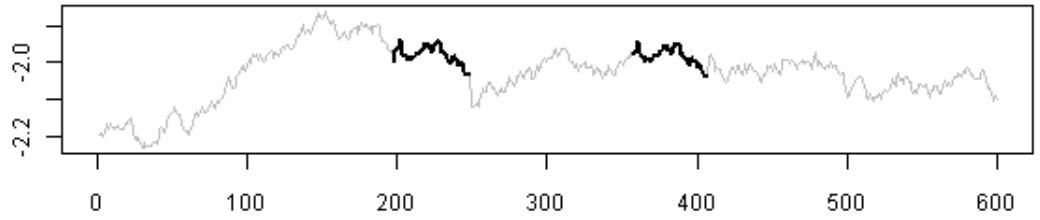


**Figura 3. Sub-série $T$ destacando a ocorrência das duas sub-seqüências $P$, após execução do algoritmo força-bruta.**

## 4. Conclusão

Esta avaliação permitiu constatar que o algoritmo de força-bruta pode ser utilizado para encontrar padrões freqüentes em series temporais com grande eficiência, em termos de precisão, apesar de sua complexidade computacional alta. No entanto, esse algoritmo é mais utilizado para fins acadêmicos uma vez que sua complexidade alta inviabiliza a sua aplicação para grande base de dados. Atualmente tenta-se desenvolver algoritmos com complexidade baixa e com alta eficiência, por exemplo, o *Time Series Projection* [Chiu et al. 2003], o qual apresenta complexidade menor caso comparado com o algoritmo de força-bruta, porém pode não identificar todos as ocorrências de um padrão freqüente.

Os testes executados possibilitaram uma melhor compreensão dos parâmetros e conceitos envolvidos na identificação de padrões freqüentes em séries temporais. Dando uma noção da dificuldade de estipular tais parâmetros.

Para trabalhos futuros está-se considerando novas formas de aumentar o desempenho do algoritmo de força-bruta, sobretudo com o uso de estruturas de indexação como M-Trees [Ciaccia et al. 1997].

## Referências

Chiu, B., Keogh, E., and Lonardi, S. (2003). Probabilistic discovery of time series motifs. In *Proceedings of the 9th International Conference on Knowledge Discovery and Data Mining*, pages 493–498, New York, USA. ACM Press.

Ciaccia, P., Patella, M., and Zezula, P. (1997). M-tree: An efficient access method for similarity search in metric spaces. In *VLDB '97: Proceedings of the 23rd International Conference on Very Large Data Bases*, pages 426–435, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Fink, E. (2004). *Data Mining in Time Series Databases*, volume 57 of *Machine perception and artificial intelligence*, chapter Indexing of Compressed Time Series, pages 43–65. World Scientific, Singapore.

Lin, J., Keogh, E., Lonardi, S., and Patel, P. (2002). Finding motifs in time series. In *Proceedings of the Second Workshop on Temporal Data Mining at the Eighth International Conference on Knowledge Discovery and Data Mining*, pages 53–68, Edmonton, Alberta, Canada. ACM Press.

Morettin, P. A. and Toloi, C. M. (2006). *Anï¿½lise de Sï¿½ries Temporais*. Edgard Blï¿½cher, Sï¿½o Paulo, Brasil, 2 edition.

Rezende, S. O. (2003). *Sistemas Inteligentes: Fundamentos e Aplicaï¿½ï¿½es*. Manole, Barueri, Brasil.

Tatavarty, G., Bhatnagar, R., and Young, B. (2007). Discovery of temporal dependencies between frequent patterns in multivariate time series. In *Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining*, pages 688–696, Honolulu, Hawaii, USA. IEEE.