Journal of the Brazilian Computer Society, 202X, XX:1, doi: 10.5753/jbcs.202X.XXXXXX

Distância de Edição em Estruturas Hierárquicas: Um Estudo entre Algoritmo Exato e Heurística Aproximada

Lucas Gualtieri [**PUC Minas** | ***lgualtieri@sga.pucminas.br***]
Gabriel Quaresma [**PUC Minas** | ***gabrie.oliveira.1462924@sga.pucminas.br***]
Luca Gonzaga [**PUC Minas** | ***lucalourenco@gmail.com***]
Pedro Alves [**PUC Minas** | ***Pedro.alves.1446100@sga.pucminas.br***]

Pontifical Catholic University of Minas Gerais, R. Dom José Gaspar, 500 - Coração Eucarístico, Belo Horizonte - MG, 30535-901.

Tree Edit Distance (TED) is a well-established metric for measuring structural similarity between trees, with applications ranging from computational biology to image analysis. This paper presents a comparative study between two approaches for calculating TED: the exact algorithm proposed by Zhang and Shasha (1989), and a heuristic method based on tree serialization followed by Levenshtein distance calculation. The study includes the implementation of both algorithms and an experimental evaluation using synthetic datasets with varying tree topologies and sizes. Results show that while the Zhang and Shasha algorithm offers precise distance computation at higher computational cost, the heuristic approach provides significant runtime improvements with acceptable approximation for large trees. The trade-offs between accuracy and performance are discussed, offering insights into the suitability of each method for different application contexts.

Keywords: Tree Edit Distance, Zhang and Shasha Algorithm, Levenshtein Distance, Approximation Heuristics, Tree Comparison, Computational Efficiency.

1 Introdução

A distância de edição entre árvores (Tree Edit Distance, TED) é uma medida clássica de similaridade estrutural entre duas árvores. Ela é definida como o número mínimo de operações (substituição, inserção ou remoção de vértices) necessárias para transformar uma árvore T1 em outra árvore T2. Essa métrica tem aplicações importantes em diversas áreas, como comparação de estruturas secundárias de RNA, árvores filogenéticas e análise hierárquica de imagens.

Neste trabalho, implementamos o algoritmo proposto por Zhang and Shasha [1989], que é um dos algoritmos mais conhecidos para cálculo exato da TED em árvores ordenadas, e o comparamos com um segundo método baseado em heurística via serialização e aplicação da distância de Levenshtein.

2 Objetivo

O objetivo deste trabalho é:

- Implementar o algoritmo exato de Zhang and Shasha [1989] para distância de edição entre árvores.
- Implementar um segundo algoritmo de comparação de árvores baseado em heurística.
- Comparar os dois métodos em relação ao tempo de execução e a diferença absoluta das TEDs (Tree Edit Distance) encontradas, conforme o tamanho das árvores aumenta.

3 Metodologia

3.1 Algoritmo de Zhang & Shasha (1989)

Este algoritmo utiliza programação dinâmica para calcular a distância de edição entre árvores ordenadas. Ele é baseado na decomposição das árvores em florestas e no reuso de resultados de subproblemas.

- Complexidade: $O(n^3)$ no pior caso.
- Entrada: duas árvores ordenadas com rótulos.
- Saída: distância de edição entre as duas árvores.

3.2 Segundo Método: Serialização + Levenshtein

Neste método alternativo, transformamos cada árvore em uma string por meio de uma travessia em pré-ordem e comparamos as duas strings utilizando a distância de Levenshtein.

- Complexidade: $O(n^2)$ para Levenshtein + custo de serialização.
- Vantagem: simples e eficiente para grandes instâncias.
- Desvantagem: perde informação estrutural da árvore.

3.3 Geração de Dados

Utilizamos um gerador de árvores com diferentes topologias:

- Árvores binárias completas
- Árvores lineares
- Árvores estrela
- Árvores aleatórias

- Árvores rasas (de baixa altura)

Os tamanhos variaram de 2^3 até 2^{10} nós.

4 Resultados Experimentais

4.1 Métricas Avaliadas

- Tempo de execução (em segundos)
- Distância de edição (valor absoluto)

4.2 Metodologia de Execução

- As árvores foram carregadas a partir de arquivos binários.
- Os algoritmos foram executados para todos os pares de árvores.
- Cada execução foi repetida várias vezes para cálculo da média.
- Os resultados foram salvos em arquivos CSV para análise posterior.

4.3 Resultados Obtidos

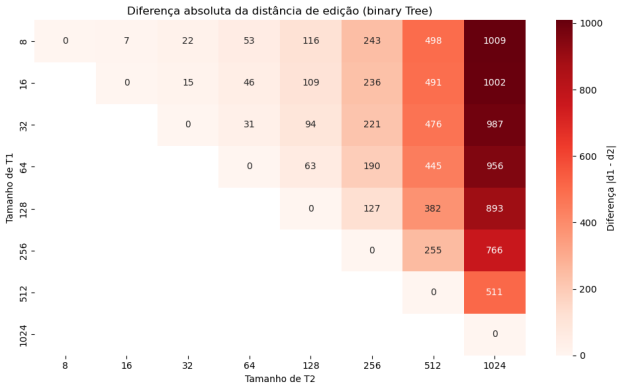

Figure 1. Diferença absoluta da distância de edição em árvores binárias.

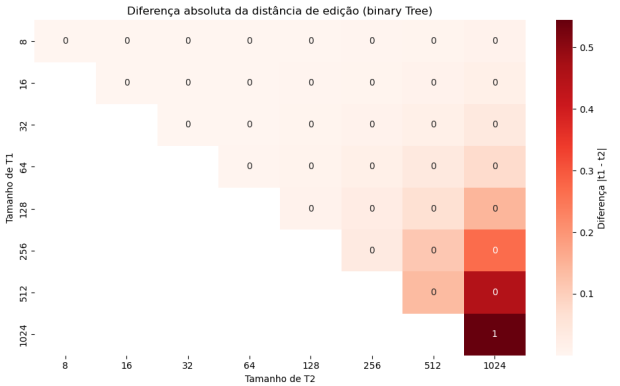

Figure 2. Diferença absoluta do tempo de execução dos algoritmos em árvores binárias.

Figure 3. Diferença absoluta da distância de edição em árvores lineares (Grafo de Linha).

Resumo dos resultados:

Figure 4. Diferença absoluta do tempo de execução dos algoritmos em árvores lineares (Grafo de Linha).

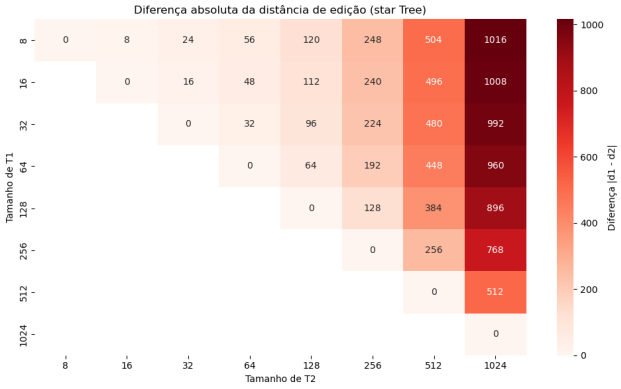

Figure 5. Diferença absoluta da distância de edição em árvores estrela.

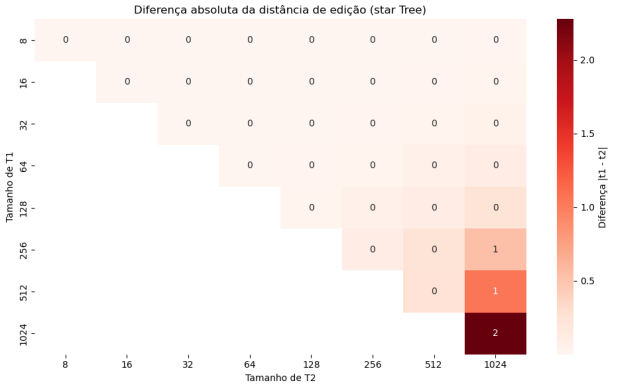

Figure 6. Diferença absoluta do tempo de execução dos algoritmos em árvores estrela.

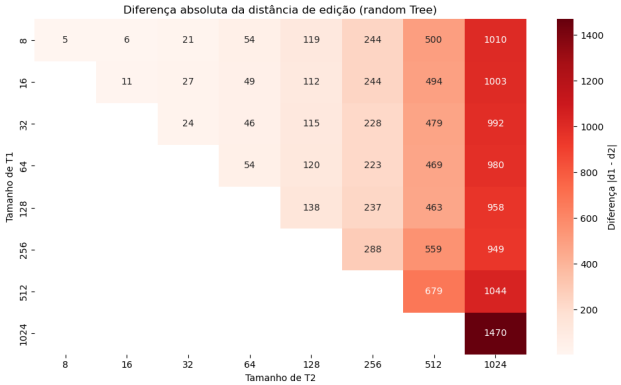

Figure 7. Diferença absoluta da distância de edição em árvores aleatórias.

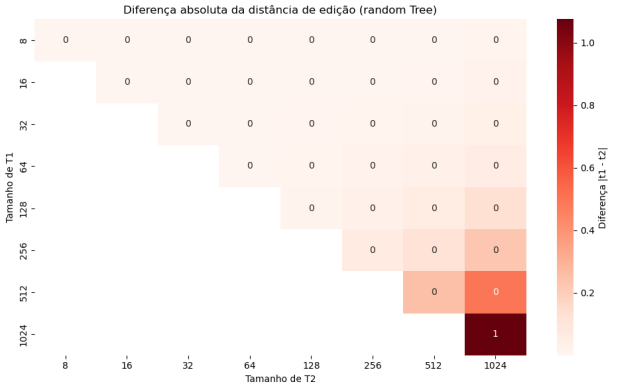

Figure 8. Diferença absoluta do tempo de execução dos algoritmos em árvores aleatórias.

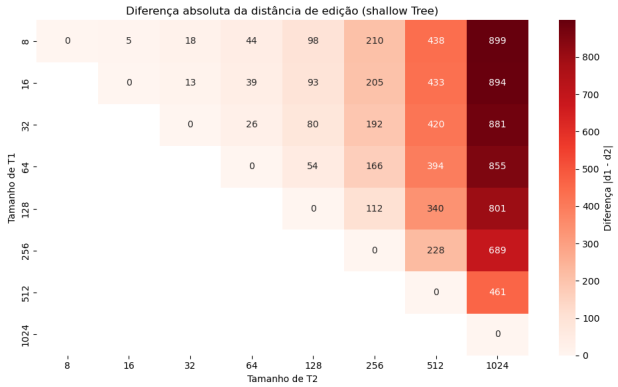

Figure 9. Diferença absoluta da distância de edição em árvores aleatórias de baixa altura.

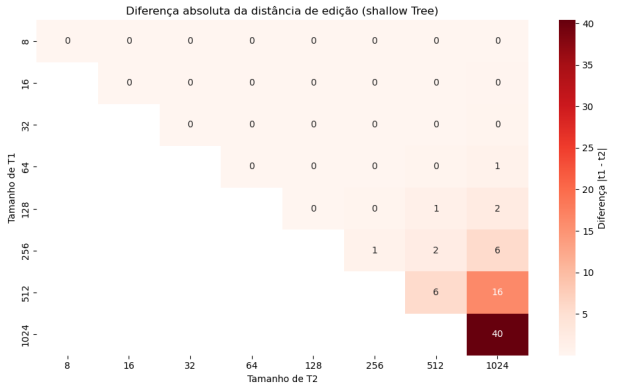

Figure 10. Diferença absoluta do tempo de execução dos algoritmos em árvores aleatórias de baixa altura.

- O algoritmo de Zhang and Shasha [1989] apresentou crescimento cúbico em tempo.
- O método heurístico foi mais rápido, mas muito menos preciso.

5 Discussão

Os resultados confirmam as expectativas teóricas: o algoritmo de Zhang and Shasha [1989] é ótimo, portanto sempre encontra a menor TED, mas seu custo pode se tornar impraticável para certas aplicações. A abordagem heurística é útil quando a eficiência é mais importante do que a otimalidade. Os heatmaps revelam que o algoritmo heurístico, apesar de seu excelente desempenho em tempo, produz estimativas de distância significativamente inferiores àquelas fornecidas pelo algoritmo de Zhang & Shasha, especialmente em árvores de grande porte e topologias mais complexas, como árvores aleatórias. A discrepância ultrapassa 1000 unidades em diversos casos, o que compromete a aplicabilidade do método em contextos sensíveis à exatidão da métrica. Em contrapartida, a eficiência temporal do método heurístico o torna útil em aplicações com restrições de tempo ou com tolerância a aproximações.

6 Conclusão

Este trabalho demonstrou, de forma experimental, as diferenças entre dois métodos para calcular distância de edição entre árvores. O algoritmo de Zhang and Shasha [1989] continua sendo uma referência para comparação exata, enquanto métodos heurísticos podem ser utilizados em cenários que exigem desempenho.

Contributions

As contribuições de cada membros foram como mostrado à seguir:
Lucas Gualtieri: Implementou o algoritmo de Zhang & Sasha e estruturou os experimentos.
Gabriel Quaresma: Implementou o Algoritmo de Serialização + Levenshtein e estruturou os experimentos.
Luca Gonzaga: Estruturou os experimentos e documentou o experimento.
Pedro Alves: Implementou estruturas de dados auxiliares e documentou o experimento.

References

Zhang, K. and Shasha, D. (1989). Simple fast algorithms for the editing distance between trees and related problems. *SIAM Journal on Computing*, 18(6):1245–1262. DOI: 10.1137/0218082.