



Algoritmo de A* Paralelo para Detecção de Comunidades

Rudson Francisco da Silva Mendes

Universidade do Vale do Itajaí
R. Uruguai, 458 – Centro, 88302-202 – Itajaí – SC
rudsonm@edu.univali.br

Matheus Mendes Bruzaca

Universidade do Vale do Itajaí
R. Uruguai, 458 – Centro, 88302-202 – Itajaí – SC
matheus.bruzaca@gmail.com

Rafael De Santiago

Universidade do Vale do Itajaí
R. Uruguai, 458 – Centro, 88302-202 – Itajaí – SC
rsantiago@univali.br

RESUMO

A análise de um grafo pode necessitar o conhecimento de sua estrutura. Por exemplo, compreender as funcionalidades de cada vértice pode ser uma tarefa impossível para um analista humano em grafos com milhares de vértices. Neste contexto, subgrupos de vértices fortemente relacionados são chamados de comunidades. A necessidade de detectar comunidades se deve em maior parte para fins de melhor análise de dados, destacando um subgrupo de indivíduos que compartilham comportamentos ou atributos semelhantes. Este trabalho é sobre uma heurística baseada no algoritmo de A* paralelizável para detecção de comunidades através da função de Modularidade por Densidade.

A Modularidade por Densidade define uma função objetivo para o problema de detecção de comunidades baseada na propriedade modular, é utilizada a diferença entre a conectividade interna e externa para definir quais vértices participam de cada comunidade.

Um algoritmo de A* busca uma solução em um espaço para encontrar caminhos utilizando uma função objetivo (distância) e uma função heurística. A função objetivo é utilizada para identificar quão longe a busca está de seu objetivo. A função heurística é utilizada para evitar “obstáculos” que interagem com esta distância. Devido a sua natureza, há diversas aplicações de A* para resolver problemas de *path finding* na literatura.

Dada uma solução inicial, para cada iteração do algoritmo A* desenvolvido, são geradas soluções vizinhas por uma busca local semelhante ao método de Louvain, migrando vértices de uma comunidade à outra. Todas as soluções são armazenadas em um *heap* que prioriza a função $f(s)=g(s)+h(s)$ que depende de duas funções (objetivo e heurística). A função g é a função objetivo da Modularidade por Densidade de uma solução s ; e h é definida pelo maior valor Modularidade por Densidade entre as comunidades de s .

Para otimizar o tempo de execução do algoritmo A* sequencial, utilizou-se a tecnologia Open MPI para paralelizar o algoritmo. O problema é dividido em subproblemas que são distribuídos entre n computadores. Nesta estratégia, cada computador gera sua própria solução inicial. Os computadores geram soluções vizinhas em um intervalo de vértices distintos entre si, estes intervalos são proporcionais a quantidade de vértices do grafo. Dada uma determinada frequência, cada computador compartilha para os demais a melhor solução encontrada até então. Esta tentativa demonstrou aumento na qualidade das soluções obtidas.

Foram analisadas 11 instâncias. Na versão paralela utilizando 4 computadores, obteve-se em média 98.5% dos valores ótimos de modularidade por densidade em relação à heurística sequencial. Em duas instâncias, a versão paralela demandou mais tempo que a sequencial. Desconsiderando as discrepâncias, é obtida uma aceleração de tempo médio de 5.7 vezes melhor.

PALAVRAS CHAVE. Detecção de Comunidades, Algoritmo A*, Paralelismo.