Agrupamento de itens e dados através da implementação do algoritmo de Ant-Clustering

Lucas O. Rocha[1]

[1]Departamento de Ciência da Computação – Universidade do Estado de Santa Catarina (UDESC)
Santa Catarina – SC – Brasil

lucas.rocha147@edu.udesc.br

Resumo. *Algoritmos de Ant-Clustering tem sido estudados e aprimorados há muito tempo, provando serem eficazes nos mais diversos tipos de uso, como mineração e agrupamento de dados. Este relatório propõe-se a realizar um estudo de forma geral acerca deste princípio, implementando o básico do algoritmo, realizando algumas alterações ao mesmo, e analisando os resultados obtidos após testes de execução, observando o desempenho sobre dois cenários: o agrupamento de itens homogênios, e o agrupamento de dados heterogênios.*

Abstract. *Ant-Clustering algorithms have been studied and improved for a long time, proving to be effective in the most diverse cases of use, such as data mining and data grouping. This paper proposes to accomplish a general study about this principle, implementing the basic of the algorithm, realizing some changes, and analyzing results obtained after running tests, observing the performance over two scenarios: the clustering of homogeneous items, and the clustering of heterogeneous data.*

1. Introdução

O agrupamento artificial de itens e dados baseados em colônias de formigas tem como objetivo realizar agrupamentos de forma concisa e coesa, o que se torna possível através da simulação de um ambiente semelhante ao encontrado ao se estudar o comportamento de formigas que realizam um processo de limpeza e organização de seus formigueiros, chamado Self-Organization [Bonabeau et al. 1999].

Com base nisso, iniciou-se o estudo e implementação de um algoritmo que seja capaz de reproduzir o conceito básico de Ant-Clustering, focado em duas partes: o agrupamento de itens de caráter homogêneo, e o agrupamento de dados de caráter heterogêneo. Para isso, faz-se necessário um algoritmo que seja capaz de simular: um ambiente e a criação de agentes que se movimentem pelo mesmo; a distribuição uniformemente aleatória de itens/dados pelo ambiente; e a tomada de decisão dos agentes em pegar ou largar um item/dado.

Para tal, foram utilizados como base alguns artigos já presentes na literatura, que fornecem noções e conceitos do algoritmo básico, e de versões mais experimentais do modelo de Ant-Clustering, como o escrito por [Bonabeau et al. 1999], que referencia o processo de agrupamento baseado em colônias. Também, [Lumer and Faieta 1994] apresentaram fórmulas iniciais para o processo de agrupamento, que serão utilizadas como base nesta implementação. Ainda, o proposto por [Handl et al. 2003] foi estudado e utilizado para a parametrização em certos aspectos do algoritmo.

As estratégias que compõe toda a criação e lógica por trás do algoritmo serão explicadas na Seção 2. A Seção 3 apresenta os resultados obtidos através das experimentações. Na Seção 4, discorre-se sobre a análise de tais resultados, enquanto a Seção 5 aborda sobre ideias a serem implementadas no futuro, além de concluir o relatório.

2. Metodologia de desenvolvimento

Esta seção destina-se a relatar as ideias, escolhas, e lógicas atribuídas a implementação do algoritmo, justificando cada uma delas com base na literatura e/ou em testes empíricos realizados durante o processo de criação do sistema.

Inicialmente, pode-se apontar fatores presentes tanto no agrupamento de itens quanto no agrupamento de dados. Apesar de diferentes, as lógicas de pergar e largar itens e dados possuem pontos em comum: ambas dependem da quantidade de células presente no campo de visão dos agentes. Para este cálculo, ao longo da implementação, notou-se que a melhor forma de se obter o resultado esperado se dá através da fórmula: $(2n+1)^2$, onde *n* corresponde ao raio de visão do agente. Além disso, a movimentação dos agentes segue o mesmo princípio em ambos os cenários. Vale ressaltar também que, ao final do número de iterações, os agentes que ainda estiverem carregando items/dados são programados para terminar sua rota, garantindo que a quantidade de dados final seja a mesma que a inicial. Ainda, em ambos os casos o ambiente utilizado foi programado para se assemelhar a um eplisoide, fazendo com que os agentes, ao encontrarem a borda superior do ambiente, continuem sua movimentação partindo da borda inferior. A mesma dinâmica é válida para caso o agente encontre a borda inferior e bordas laterais.

Por questões de implementação, optou-se pela realização de um processo sequencial dos agentes, e não em paralelo, tornando a simulação menos precisa em relação a vida real. Da mesma forma, a utilização de feromônios [Dorigo and Di Caro 1999] também foi descartada na implemtentação.

A subseção 2.1 é referente a modelagem do algoritmo para agrupamento de itens homegêneos, enquanto a subseção 2.2 discorre sobre o agrupamento de dados heterogêneos.

2.1. Agrupamento de itens

Para melhor simulação do ambiente, optou-se pela implementação em uma matriz de dimensões *n x n*, sendo *n* um valor fixado em 50. Em seguida, os itens a serem agrupados foram alocados de forma aleatória ao longo do ambiente, seguindo uma distriução uniforme, assim como os agentes, que foram inicializados e distribuídos de forma randômica pelo ambiente. A tomada de decisão dos agentes entre pegar ou largar itens é puramente probabilística, uma vez que nenhum parâmetro foi fixado como constante. Para pegar um item, o agente verifica seu entorno, baseado no raio de visão, fazendo a relação entre célucas ocupadas CO e células totais CT, onde CT é expressa pela fórmula:

$$CT = (2n + 1)^2 \tag{1}$$

Assim, tem-se que a probalidade do agente pegar um item (ProbP) é dada por ProbP = $\frac{CO}{CT}$, sendo esta probabilidade comparada a um valor aleatório do intervalo [0.0, 1.0]. O agente então pegará o item caso ProbP seja menor ou igual ao número aleatório. Paralelamente, a

probabilidade do agente largar um item (ProbL) se dá por ProbL = $\frac{CO}{CT}$, porém o agente de fato largará o item caso ProbL^2 seja maior ou igual que um valor gerado aleatóriamente entre [0.0, 1.0]. A inclusão da elevação de ProbL ao quadrado, ProbL^2, tem como objetivo garantir que o agente largará um item apenas em um cenário extremente propício para isso.

2.2. Agrupamento de dados

Tendo-se implementado o algoritmo básico para agrupamento de itens homogêneos, viu-se a necessidade de adaptar certos processos e parâmetros para a realização do agrupamento com dados.

A matriz base do ambiente passou de dimensões *n x n* para $\sqrt{10N\_dados}$ x $\sqrt{10N\_dados}$, onde *N_dados* corresponde a quantidade de dados a serem agrupados, como proposto em [Handl et al. 2003]. O número de iterações foi alterado para iterações=$2000.N\_dados$, como também expresso por [Handl et al. 2003]. A distribuição dos dados se deu de forma semelhante a dos itens, porém, diferentemente do itens unidimensionais, os dados podem ser n-dimensionais. Para a tomada de decisão dos agentes, requeriu-se uma abordagem diferente da presente no agrupamento de itens. A probabilidade de se pegar um dado (ProbP) foi expressa por:

$$ProbP = 1 - sigmoid(f(o_i)) \tag{2}$$

enquanto a probabilidade de se largar um dado (ProbL) deu-se por:

$$ProbL = sigmoid(f(o_i)) \tag{3}$$

sendo a função $sigmoid$:

$$Sigmoid(x) = \frac{1 - e^{-cx}}{1 + e^{-cx}} \tag{4}$$

onde c é dado por $10.raio\_de\_visao$. A função $sigmoid$ faz-se necessária para realizar o ajuste da função de similaridade (5) em valores probabilísticos, como expresso por [Gao 2016]. A função de similaridade média [Lumer and Faieta 1994] é dada por:

$$f(o_i) = \frac{1}{s^2} \sum_j (1 - \frac{d(o_i, o_j)}{\alpha}), sef > 0 \tag{5}$$

Caso f seja menor ou igual a zero, a função retornará o valor 0. O valor de $d(o_i, o_j)$ é referente a distância euclidiana entre o dado da posição atual do agente e seus vizinhos, sendo expressa por:

$$d(o_i, o_j) = \sqrt{\sum_{k=1}^{m} (o_{ik} - o_{jk})^2} \tag{6}$$

O cálculo da constante α, presente na equação (5) é dado pelo somatório das distâncias euclidianas entre os dados, dividido pela quantidade de dados elavado ao quadrado.

3. Experimentações e Resultados

Neste ponto, serão descritos os resultados obtidos através das experimentações com variações de parâmetros, como o raio de visão, e a alteração entre itens e dados.

3.1. Itens

Para o agrupamento de itens, foram utilizados para os experimentos, além dos já citados na subseção 2.1, os seguintes parâmetros: raio de visão = 1 e 5; número de itens = 1000; número de agentes = 20; iterações = 1000000, item = valor inteiro qualquer.

Abaixo, estão dispostos o início da execução e o resultado final, obtidos através da experimentação de agentes com raio de visão igual a 1.

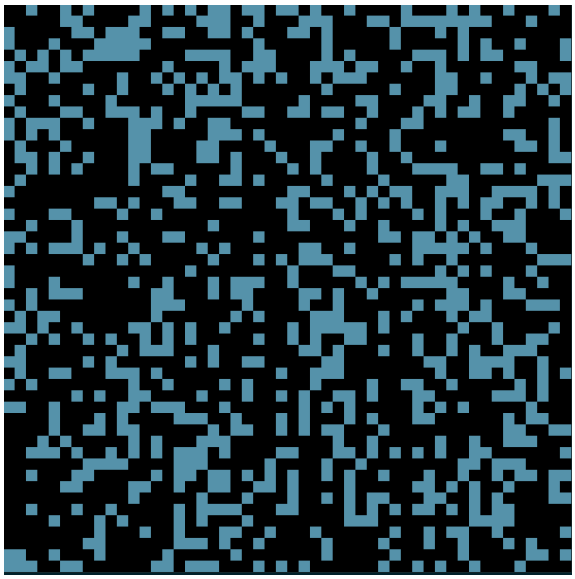

Figura 1. Início da execução - Raio de visão = 1

Figura 2. Final da execução - Raio de visão = 1

A seguir, encontra-se o início da execução e o resultado obtido experimentado com raio de visão igual a 5.

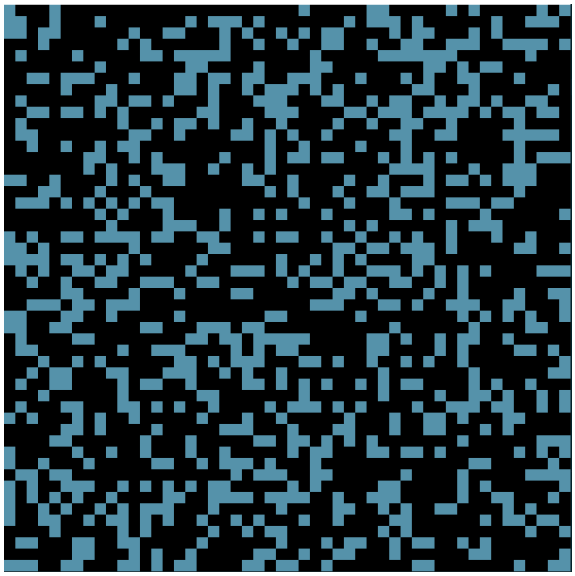

Figura 3. Início da execução - Raio de visão = 5

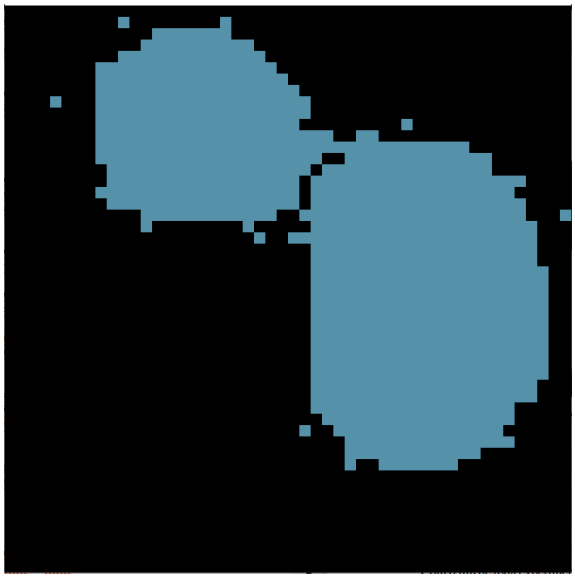

Figura 4. Final da execução - Raio de visão = 5

3.2. Dados

Para o agrupamento de dados, os únicos parâmetros não calculávies após o início da execução foram tais: raio de visão = 1; número de agentes = 20. Quanto aos dados que serão agrupados, tem-se um conjunto de 400 dados bidimensionais separados em 4 grupos e um conjunto de 600 dados bidimensionais divididos em 15 grupos. Ambas as bases de dados foram geradas de forma aleatória.

Para a base de dados bidimensionais com 4 grupos, a geração aleatoria se deu com os seguintes parâmetros:

- grupo 1: dados pertencentes aos intervalos [-20,2] e [-20,2].
- grupo 2: dados pertencentes aos intervalos [20,2] e [20,2].
- grupo 3: dados pertencentes aos intervalos [-20,2] e [20,2].

- grupo 4: dados pertencentes aos intervalos [20,2] e [-20,2].

originando 400 dados únicos. O parâmetro α para esta base foi calculado como: 35.15397017803196.

Logo abaixo, está o início da execução e o resultado obtido para tal experimentação:

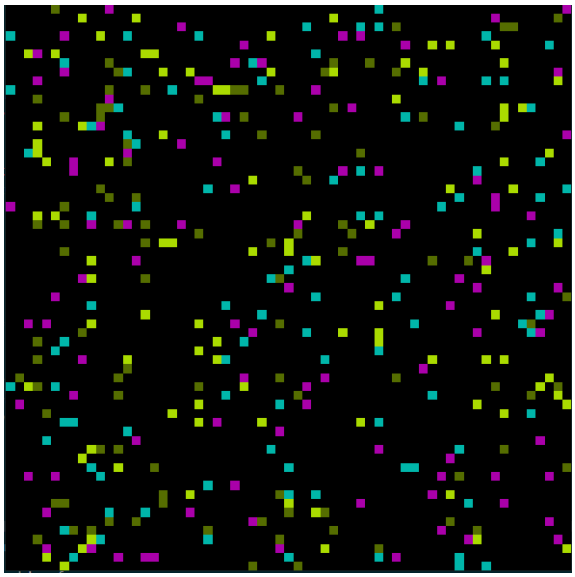

Figura 5. Início da execução - Raio de visão = 1. 4 grupos

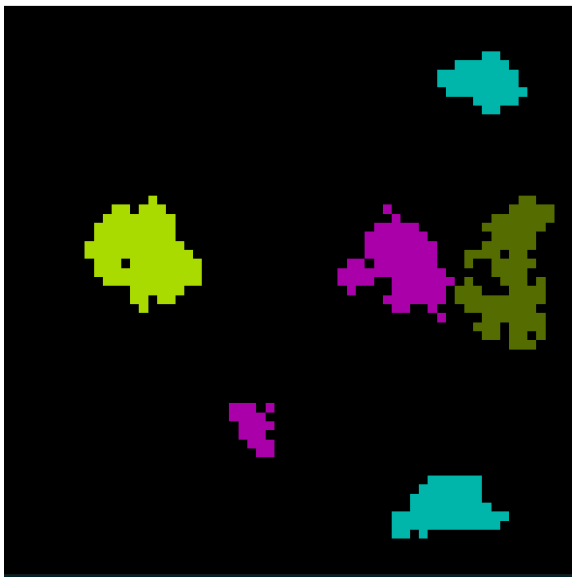

Figura 6. Final da execução - Raio de visão = 1. 4 grupos

Para a base de dados bidimensionais de 15 grupos, foi-se gerada de forma aleatória dados pertencentes ao intervalo [3.178, 17.124], originando 600 dados únicos. O parâmetro α para esta base foi calculado como: 5.745796826725059.

Abaixo, encontra-se a distribuição inicial e o agrupamento final desta base de dados.

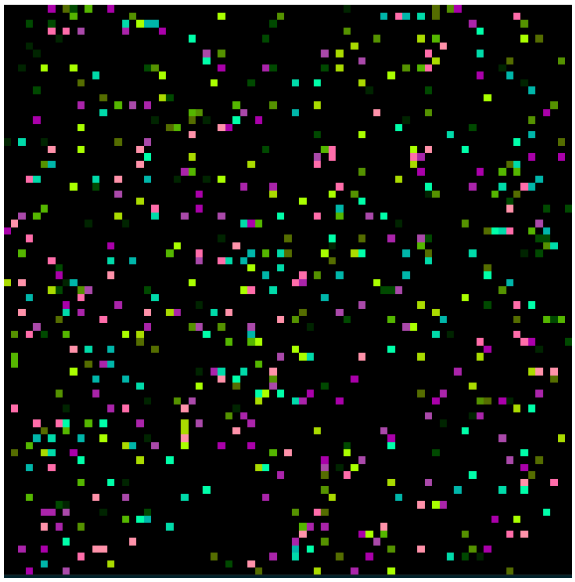

Figura 7. Início da execução - Raio de visão = 1. 15 grupos

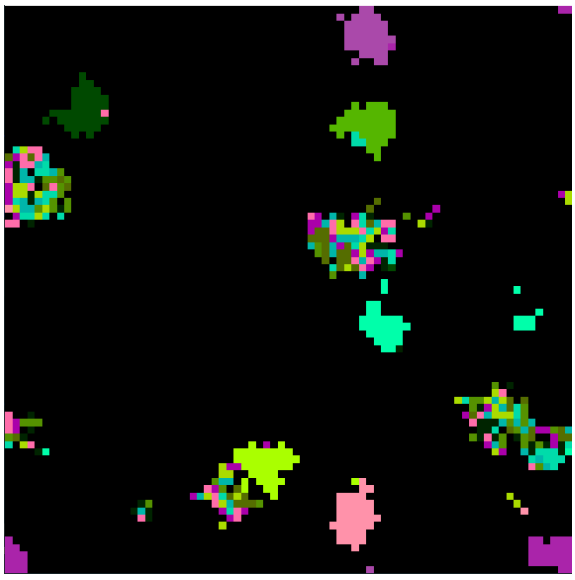

Figura 8. Final da execução - Raio de visão = 1. 15 grupos

4. Análise dos resultados

Após a execução dos processos, e ao longo de toda implementação, pode-se observar fatores que são diretamente derivados dos parâmetros e escolhas de lógicas utilizada na construção do modelo.

Analisando as figuras 2 e 4, presentes na subseção 3.1, pode-se perceber a notória diferença presente devido ao raio de visão do agente. Quanto maior for a percepção que o agente tiver, melhor será seu rendimento. Consequentemente, o número de blocos de itens tenderá a diminuir, além de ficarem mais concisos e menos esparços. Notou-se também que com raio de visão em 1, o algoritmo tende a convergir para um estado semelhante ao de raio de visão 5. Entre tanto, para isso ocorrer, o número de iterações precisaria ser muito maior do que o estabelecido no experimento. Outro ponto observado foi o tempo

de execução: para o processo com raio 1, o tempo médio foi de 4 minutos, enquanto pra raio 5, a estimativa é de que se tenham passados 7 minutos. Excluindo a capacidade e velocidade de processamento da máquina utilizada para as simulações, pode-se associar o aumento do tempo de execução com o raio de visão do agente: com um raio maior, o agente fica mais seletivo quanto a ação de pergar/largar um item, fazendo com que cada iteração demore mais para se encerrar, acarretando na adição do tempo final.

Partindo para o agrupamento de dados, e com base nas imagens 6 e 8, fica evidente que a semelhança entre os grupos da base de dados com 15 grupos teve papel fundamental para um desempenho questionável do algoritmo. Enquanto que na figura 6 pode-se perceber com clareza a formação dos grupos com dados semelhantes, na figura 8, ainda que estes também estejam presentes, há aglutinações de dados pertencentes a grupos diferentes. Isto se deve a fatores como a constante α, que é consideravelmente menor para esta base do que para de 4 grupos. Isto pois, diferentemente da clara divisão entre os 4 grupos da base de 400 dados, a base com 600 dados apresenta uma diversificação muito menor entre os mesmos, como pode ser visto analisando os dados [10.256, 9.251] e [11.988, 9.926], que pertencem a dois grupos diferentes. Além disso, e assim como na experimentação de itens, notou-se um aumento do médio do tempo de execução de acordo com a crescente dos grupos, sendo de 8 muinutos para a base com 4 grupos e 10 minutos para a base de 15 grupos.

5. Conclusão e Trabalhos futuros

Com o fim de toda experimentação, fica notória a capacidade humana de ao menos simular com certa precisão ambientes reais. Este estudo é completamente baseado em organismos reais, que apesar de simples, conseguem se auto-organizar de forma exemplar. Simular este processo mostrou ser de grande valia.

Ainda, percebe-se que a prática de testes e estudos ao longo do desenvolvimento é fundamental para o conhecimento científico, uma vez que muitas das tomadas de decisão ocorreram após a verificação de desempenho, feita através de testes. Obviamente, a literatura já estabelecida se mostra completamente necessária para ciência visto que é a principal base para o conhecimento, tanto científico, quanto humano.

Por fim, para o futuro, espera-se aprimorar o agrupamento de dados, aperfeiçoando ou até criando uma nova metodologia na tomada de decisão dos agentes. Entra também nos horizontes da pesquisa, a implementação de um modelo muitos mais próximo a realidade, que possa não só simular, mas copiar por completo o comportamento do seres tomados como base para este estudo.

Referências

Bonabeau, E., Theraulaz, G., Dorigo, M., Theraulaz, et al. (1999). *Swarm intelligence: from natural to artificial systems*. Number 1. Oxford university press.

Dorigo, M. and Di Caro, G. (1999). Ant colony optimization: a new meta-heuristic. In *Proceedings of the 1999 congress on evolutionary computation-CEC99 (Cat. No. 99TH8406)*, volume 2, pages 1470–1477. IEEE.

Gao, W. (2016). Improved ant colony clustering algorithm and its performance study. *Computational Intelligence and Neuroscience*, 2016.

Handl, J., Knowles, J., and Dorigo, M. (2003). Ant-based clustering: a comparative study of its relative performance with respect to k-means, average link and 1d-som. In *Proceedings of the Third International Conference on Hybrid Intelligent Systems, IOS Press*.

Lumer, E. D. and Faieta, B. (1994). Diversity and adaptation in populations of clustering ants. In *Proceedings of the third international conference on Simulation of adaptive behavior: from animals to animats 3: from animals to animats 3*, pages 501–508.