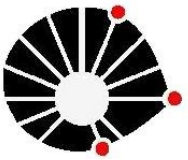

Universidade Estadual de Campinas
Faculdade de Ciências Aplicadas

Projeto de Pesquisa - Iniciação Científica

Algoritmos de agrupamento e de detecção de comunidades em redes: comparações e aplicações.

Aluno: Vitor Alves Iglesias
Orientadora: Profa. Dra. Priscila Cristina Berbert Rampazzo

Maio de 2024.

Resumo

O agrupamento é uma das técnicas de Aprendizado de Máquina mais importantes, com aplicações práticas em diversas áreas de conhecimento. Neste contexto, uma abordagem vem ganhando notoriedade em trabalhos recentes: os algoritmos de agrupamento baseados em redes. Estes algoritmos constroem uma rede a partir dos dados e utilizam detecção de comunidades para encontrar os grupos. De acordo com a aplicação, estes grupos podem nos trazer importantes informações. Entender os métodos desta classe de agrupamentos e suas características é de extrema importância para realizar um agrupamento bem sucedido. A primeira etapa deste projeto tem como objetivo realizar uma análise comparativa entre algoritmos de agrupamento e algoritmos de detecção de comunidades. Na segunda etapa, serão criadas bases de dados para análises de problemas específicos e reais, como identificação de comunidades e grupos de pesquisa em universidades, identificação de comunidades e linhas de pesquisa em revisões bibliográficas, etc. Todos os algoritmos serão implementados em linguagem de programação Python.

Palavras-chave: Agrupamento, Detecção de Comunidades, Redes.

1 Introdução

Os avanços computacionais e tecnológicos fizeram crescer a obtenção de informações. Isso resultou em um crescimento exponencial da quantidade de dados gerados e armazenados. Consequentemente, surge um grande desafio: como extrair e processar informações relevantes deste vasto conjunto de dados.

Em Aprendizado de Máquina, as técnicas de agrupamento, do inglês *Clustering*, tem como objetivo particionar a base de dados em conjuntos de classes (*clusters*). Essa divisão é realizada a partir de critérios de agrupamento, que verificam quão similar ou dissimilar um dado é quando comparado a outro. Dependendo da escolha desses critérios, encontramos diferentes formas de agrupar o conjunto de dados. Quando dados são agrupados, a proximidade é usualmente indicada através de medidas que indicam similaridades ou dissimilaridades. Em relação à similaridade, quanto maior o valor observado, mais parecido são os dados (exemplo: coeficiente de correlação). Em relação à dissimilaridade, quanto maior o valor observado, menos parecidos os dados são (exemplo: distâncias). Os métodos clássicos de clusterização de dividem em quatro grupos: agrupamento hierárquico, agrupamento baseado em centróide, agrupamento baseado em grafo e agrupamento base-

ado em densidade (XU; TIAN, 2015). A Figura 1 ilustra exemplos de agrupamentos de dados obtidos por algoritmos clássicos. Cada algoritmo tem seu desempenho dependente no tipo de base de dados, tipicamente dependente da dimensão do espaço vetorial no qual os pontos estão distribuídos, número de amostras e formato dos grupos.

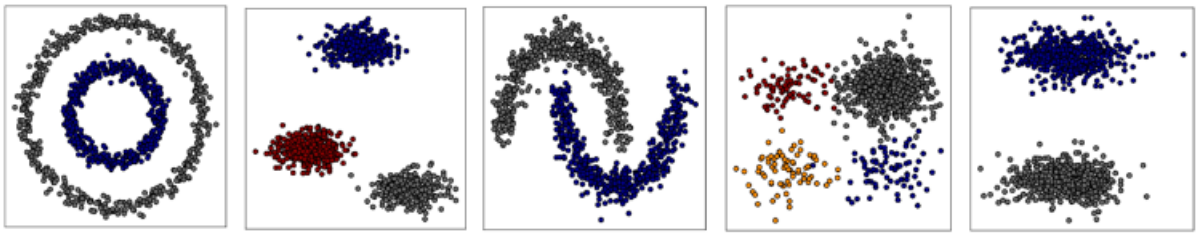

Figura 1: Exemplos de agrupamentos de dados (Fonte: (scikit-learn, 2024)).

As técnicas de agrupamento de dados baseadas em rede (SILVA et al., 2019) têm se tornado cada vez mais populares. A literatura da área tem produzido resultados significativos em comparação com algoritmos clássicos de agrupamento. Redes são estruturas relacionais definidas por entidades que se integram e interagem dinamicamente entre si (TEIXEIRA et al., 2008). Nas redes, as entidades são representadas por vértices e seus relacionamentos por arestas (Figura 1). Em algoritmos de agrupamento baseados em rede, o primeiro passo é transformar o conjunto de dados em uma rede. Em seguida, um algoritmo de detecção de comunidade é usado para particionar os vértices da rede em comunidades. Finalmente, as comunidades são interpretadas como grupos no conjunto de dados original.

Um dos benefícios da utilização de algoritmos de detecção de comunidade é que eles podem examinar a estrutura topológica que surge da transformação dos dados em uma rede, além dos atributos dos objetos no conjunto de dados (QUEROBIM, 2021).

Compreender a estrutura de comunidades em uma rede fornece percepções valiosas sobre a organização e o funcionamento dos sistemas complexos. Essas comunidades podem representar grupos de indivíduos com interesses comuns em redes sociais, grupos de documentos em redes de coautoria, linhas de pesquisa em grupos de pesquisadores, etc. Identificar comunidades permite entender melhor a dinâmica e os padrões de interação entre os elementos da rede, facilitando a análise de sua estabilidade, propagação de informações e até mesmo a identificação de potenciais pontos de influência. Assim, a detecção de comunidades desempenha um papel crucial na compreensão dos sistemas complexos e na formulação de estratégias eficazes em uma variedade de domínios.

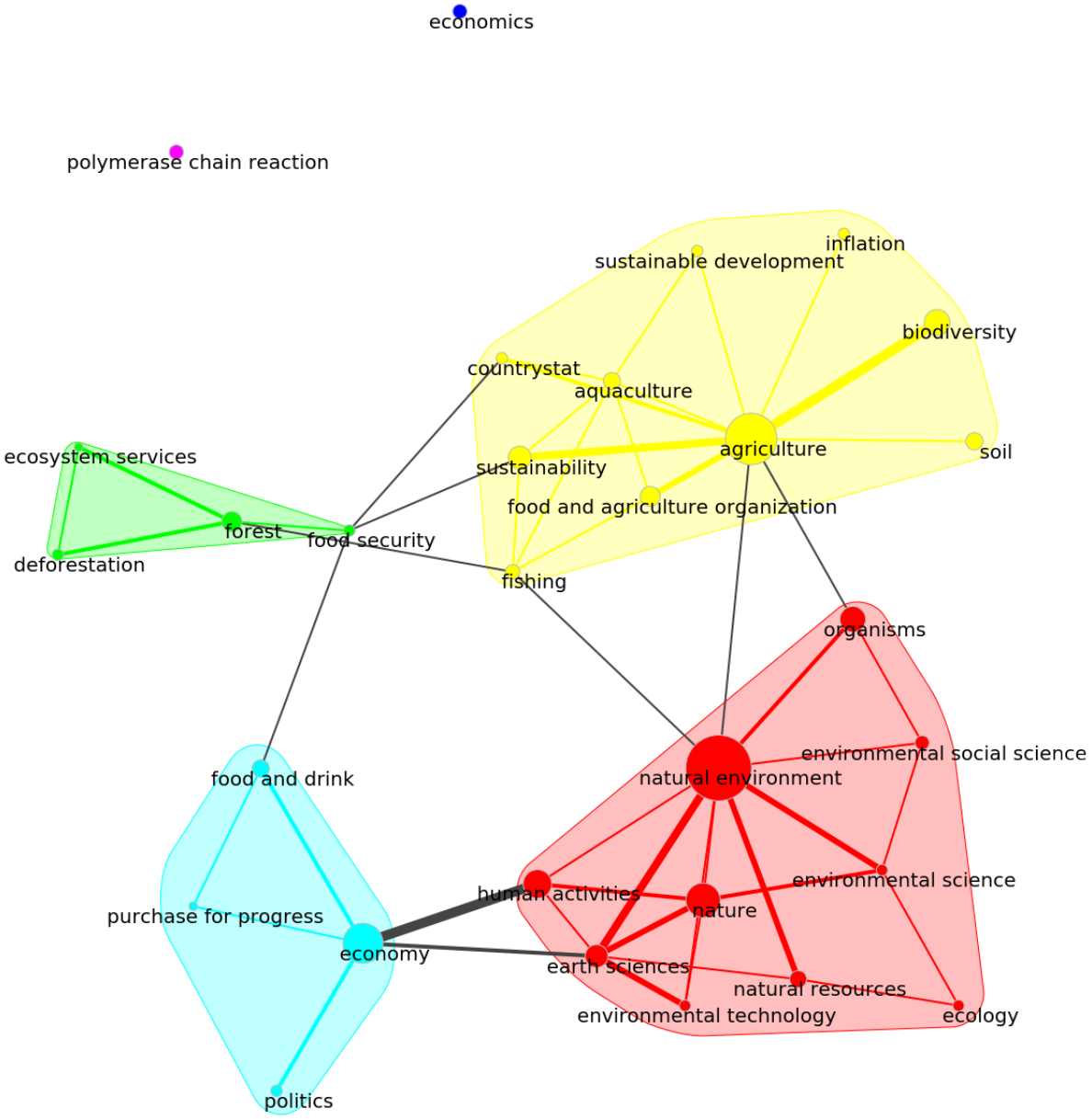

Figura 2: Exemplo de agrupamento de dados baseadas em rede (autoria própria).

2 Objetivos e Relevância da Pesquisa

Este projeto tem o intuito de explorar algoritmos de agrupamento e de detecção de comunidades em redes com enfoque multidisciplinar. A detecção de comunidades é de grande importância em sociologia, biologia e ciência da computação, disciplinas onde os sistemas são frequentemente representados como gráficos. Este problema é muito difícil e ainda não foi resolvido de forma satisfatória, apesar do enorme esforço de uma grande comunidade interdisciplinar de cientistas que trabalharam nele nos últimos anos (FORTUNATO, 2010).

O objetivo geral é realizar uma análise e comparações entre algoritmos de detecção de comunidades e aplicá-los em situações de interesse para realização de análises e obtenção de informações. Os objetivos específicos podem ser detalhados como:

1. Exploração dos algoritmos de agrupamento e detecção de comunidades.

2. Coletar dados e criar redes para aplicações reais: como identificação de comunidades e grupos de pesquisa em universidades, identificação de comunidades e linhas de pesquisa em revisões bibliográficas.

3. Analisar as informações obtidas através da detecção de comunidades nas redes das aplicações.

3 Metodologia

Diversos algoritmos podem ser utilizados, os quais possuem estratégias diferentes e podem ser aplicados em redes com características topológicas diferentes. Esta pesquisa iniciará com uma Revisão Bibliográfica sobre algoritmos de agrupamento e detecção de comunidades. Para entender o funcionamento destes algoritmos, experimentos em dados artificiais serão executados.

Algoritmos de agrupamento baseados em redes requerem a construção de uma rede a partir dos dados obtidos. Na segunda etapa da pesquisa, o serão implementados algoritmos para possibilitar a automatização da coleta de dados para criação de redes para aplicações reais: como identificação de comunidades e grupos de pesquisa em universidades, identificação de comunidades e linhas de pesquisa em revisões bibliográficas. Essa automatização incluirá a leitura automática de termos em arquivos.

Com estas novas redes, considerando aplicações reais, os algoritmos considerados na primeira etapa da pesquisa serão reavaliados. Comparações e discussões sobre as informa-

ções obtidas serão realizadas. As implementações computacionais serão em linguagem de programação Python.

3.1 Algoritmos de Detecção de Comunidades

O estudo de diferentes algoritmos de detecção de comunidades é importante por várias razões (NEWMAN; GIRVAN, 2004). Cada rede pode ter características únicas, como tamanho, densidade, distribuição de conexões e padrões de comunidade; é importante identificar o algoritmo mais adequado para cada tipo de rede. Algoritmos de detecção de comunidades podem se basear em diferentes princípios e técnicas, como otimização de modularidade, propriedades estruturais, métodos de aprendizado de máquina, entre outros. Explorar uma variedade de abordagens permite entender melhor as vantagens e limitações de cada uma e entender como interpretar as informações resultantes dos agrupamentos. Dentre os algoritmos que serão abordados nesta pesquisa, podemos destacar:

- Método de Girvan-Newman: este algoritmo identifica comunidades em uma rede removendo recursivamente as arestas com maior valor de *betweenness centrality*: uma medida de centralidade que indica a frequência com que um nó é atravessado pelos caminhos mais curtos entre todos os pares de nós na rede. Para calcular a *betweenness centrality* de um nó, é necessário determinar todos os caminhos mais curtos entre todos os pares de nós na rede e, em seguida, contar quantas vezes o nó em questão está presente em um desses caminhos. Quanto mais vezes um nó estiver presente nos caminhos mais curtos, maior será sua *betweenness centrality*. Esta medida é importante em muitos contextos, incluindo análise de redes sociais, transporte, redes de comunicação e estudos sobre dinâmicas de redes. Ela pode revelar nós que desempenham um papel crucial na comunicação ou fluxo de informações em uma rede, mesmo que esses nós não tenham muitas conexões diretas.

- Método Louvain (TRAAG et al., 2019): é um algoritmo de otimização que maximiza uma medida de modularidade para encontrar comunidades em redes. Ele funciona reatribuindo nós a comunidades de forma iterativa para melhorar a modularidade.

- Algoritmo de Label Propagation: neste algoritmo, os nós de uma rede são inicialmente rotulados com identificadores de comunidades e, em seguida, esses rótulos se propagam na rede com base na vizinhança dos nós.

- Algoritmo de Detecção de Propriedades de Grafos: este algoritmo identifica comunidades com base em propriedades estruturais específicas da rede, como densidade de conexões ou padrões de conectividade.

- Método de Detecção de Caminho Aleatório: Este método simula caminhadas aleatórias na rede e agrupa nós que são frequentemente visitados em um mesmo grupo. O método baseia-se na ideia de que nós pertencentes a uma mesma comunidade terão maior probabilidade de serem visitados durante a simulação de caminhadas aleatórias.

- Algoritmo de Leiden (TRAAG et al., 2019): é uma técnica relativamente recente e eficaz para detecção de comunidades em redes. Ele é uma variação do algoritmo Louvain, projetado para melhorar sua escalabilidade e precisão. O algoritmo de Leiden aplica uma técnica de otimização visando maximizar a modularidade da rede e agrupar seus nós. O conceito de modularidade foi proposto para medir a força da rede através da ligação de cada nó com seus nós vizinhos, dividindo-a em módulos denominados grupos, clusters ou comunidades. Para a construção desses grupos, utilizamos o pacote `leidenalg`, com funcionamento dependente do pacote `igraph`. A principal função do pacote leidenalg é `find_partition`, que é responsável por encontrar a partição ideal. O algoritmo de Leiden tornou-se popular devido à sua eficácia em detectar comunidades em redes grandes e complexas, com milhões de nós e arestas. Ele é amplamente utilizado em diversas aplicações, incluindo análise de redes sociais, biologia de sistemas, ciência da computação e muitas outras áreas onde a estrutura de comunidades em redes é relevante.

Estes algoritmos serão abordados nesta pesquisa através de pacotes específicos que os implementam, em linguagem de programação Python.

3.2 Aplicações

Esta etapa da pesquisa contemplará uma abordagem predominantemente quantitativa de caráter descritivo-analítico embasada pela pesquisa documental atrelada implementação de algoritmos de detecção de comunidades e técnicas de análise de dados para apresentação de informações.

Serão realizadas busca e extração de bases de dados no formato csv. Serão coletadas informações sobre grupos de pesquisa, artigos acadêmicos, documentos políticos, dentre outros. Estas bases passarão por uma preparação inicial, em linguagem de programação Python, com a utilização das bibliotecas `pandas`, `plotly` e `igraph`. Exemplo: palavras-chave de um grupo de artigos podem ser listadas e contabilizadas, formando uma lista de ocorrências. Uma estrutura em rede pode ser definida; esta estrutura representa cada palavra-chave com um nó e o tamanho associado a cada nó é proporcional ao número

de ocorrências da palavra-chave na lista. Palavras-chave que aparecem juntas em um mesmo documento são ligadas por arestas; cada aresta tem um peso associado à frequência com a qual as duas palavras-chave associadas são citadas juntas, considerando todos os documentos.

Depois, serão utilizados algoritmos de detecção de comunidades para agrupar as palavras-chave e identificar *clusters* da rede. A partir desses *clusters*, podemos identificar linhas que pesquisa estabelecidas entre o conjunto de documentos considerados, além de entender as relações entre as diferentes linhas.

Dentre as possibilidades de aplicação, podemos destacar:

- Análise de Coautoria: Em redes de coautoria científica, os algoritmos de detecção de comunidades podem ser aplicados para identificar grupos de autores que colaboram frequentemente em áreas de pesquisa semelhantes, facilitando a análise de tendências e identificação de potenciais colaboradores.

- Análise de Texto: Em redes de co-ocorrência de palavras em textos, os algoritmos de detecção de comunidades podem ser aplicados para identificar tópicos ou temas semânticos distintos, facilitando a organização e categorização de grandes volumes de documentos.

- Redes Sociais: Em plataformas de redes sociais, os algoritmos de detecção de comunidades podem ser usados para identificar grupos de usuários com interesses semelhantes, facilitando recomendações de amizade, sugestões de conexões profissionais e segmentação de mercado.

A visualização (ECK; WALTMAN, 2014) é uma abordagem importante para analisar redes de relações de citação entre publicações, redes de relações de coautoria entre pesquisadores ou redes de relações de co-ocorrência entre palavras-chave. A linguagem de programação Python é rica em recursos e pacotes de visualização.

4 Cronograma

O cronograma das atividades está apresentado na Tabela 1.

O cronograma de execução de 12 meses compreende as seguintes etapas:

1. Revisão Bibliográfica sobre algoritmos de agrupamento e detecção de comunidades.

2. Análises e comparações de algoritmos de agrupamento e detecção de comunidades, em linguagem de programação Python, considerando base de dados artificiais.

Atividades	1	2	3	4	5	6	7	8	9	10	11	12
(1)												
(2)												
(3)												
(4)												
(5)												
(6)												
(7)												

Tabela 1: Cronograma do Projeto de Pesquisa.

3. Implementações computacionais para coleta e tratamento de dados para aplicações reais.

4. Construção das redes para aplicações reais.

5. Análises e comparações de algoritmos de agrupamento e detecção de comunidades, em linguagem de programação Python, considerando redes e bases de dados das aplicações reais.

6. Relatório Parcial das Atividades.

7. Relatório Final das atividades.

Está prevista também, a divulgação do trabalho de Iniciação Científica por meio de apresentação no Congresso de Iniciação Científica da UNICAMP e congressos relacionados às áreas deste projeto (SBPO 2025, ENEGEP 2025, dentre outros).

Referências

ECK, N. J. V.; WALTMAN, L. Visualizing bibliometric networks. In: *Measuring scholarly impact: Methods and practice*. [S.l.]: Springer, 2014. p. 285–320.

FORTUNATO, S. Community detection in graphs. *Physics reports*, Elsevier, v. 486, n. 3-5, p. 75–174, 2010.

NEWMAN, M. E.; GIRVAN, M. Finding and evaluating community structure in networks. *Physical review E*, APS, v. 69, n. 2, p. 026113, 2004.

QUEROBIM, J. N. C. Análise comparativa entre algoritmos de agrupamento e de detecção de comunidades em redes. Universidade Federal de São Carlos, 2021.

scikit-learn. *Clustering*. 2024. Https://scikit-learn.org/stable/modules/clustering.html.

SILVA, D. M. da; BRITO, J. A. de M.; OLIVEIRA, C. S. Um estudo computacional comparativo entre algoritmos de agrupamento e de detecção de comunidades. *Simpósio de Pesquisa Operacional e Logística da Marinha*, 2019.

TEIXEIRA, L.; LIMA, L.; ABREU, N. Grafos que modelam redes confiáveis. *Mestrado em Engenharia de Produção, COPPE, Universidade do Rio de Janeiro, Rio de Janeiro*, 2008.

TRAAG, V. A.; WALTMAN, L.; ECK, N. J. V. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, Nature Publishing Group UK London, v. 9, n. 1, p. 5233, 2019.

XU, D.; TIAN, Y. A comprehensive survey of clustering algorithms. *Annals of Data Science*, Springer, v. 2, n. 2, p. 165–193, 2015.