Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 8, n. 1, 2021.

Trabalho apresentado no XL CNMAC, Evento Virtual - Co-organizado pela Universidade do Mato Grosso do Sul (UFMS).

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Agrupamento de padrões de caminhos em treinamento virtual: uma análise de similaridades

Alexandre Pereira de Faria[1]
PPGMNE/UFPR, Curitiba, PR
Klaus de Geus[2]
PPGMNE/UFPR, Curitiba, PR
Sérgio Scheer[3]
PPGMNE/UFPR, Curitiba, PR

Resumo. Em sistemas virtuais de treinamento os mecanismos de rastreamento das interações dos usuários geram um conjunto de dados a partir dos quais é possível avaliar a execução das tarefas instrucionais e inferir o estado de conhecimento do aprendiz. A representação das tarefas como grafos permite o mapeamento da sua ordem de execução e comparação com a uma sequência esperada. Este trabalho tem como objetivo analisar os métodos de agrupamentos destes grafos a partir de dados de interação em um sistema virtual de treinamento profissional para atividades de linha viva em subestações elétricas. Foram realizados estudos com diferentes definições de similaridade de grafos. Ao final são apontadas aquelas com as quais se obtiveram os melhores agrupamentos com destaque para aquelas derivadas das distâncias matriciais.

Palavras-chave. Análise de Agrupamento, Similaridade de Grafos, Mineração de Dados Educacionais

1 Introdução

Sistemas Tutores Inteligentes, Jogos Sérios e Simulações fazem parte de uma modalidade de instrução caracterizada pela interatividade, contextualização, devolutiva automática dos resultados da aprendizagem e adaptação ás necessidades do aprendiz [4]. Dois desafios fundamentais no desenvolvimento destes sistemas são a avaliação da aprendizagem e a visualização da performance do aprendiz. A automação de procedimentos instrucionais, como a avaliação, dependem do rastreamento e registro das interações dos usuários na forma de um conjunto de dados que represente o estado de conhecimento do usuário ou modelo do aprendiz. As técnicas de mineração de dados educacionais tem contribuído tanto para a categorização quanto para a inferência do modelo do aprendiz. Dentre estas técnicas a extração de padrões similares tem possibilitado a categorização dos aprendizes segundo seu estado de conhecimento. A identificação de classes de aprendizes define níveis de execução de uma tarefa e permite a criação de sistemas de recomendação tornando o processo de aprendizagem adaptável as necessidades do aprendiz.

Uma tarefa instrucional pode ser mapeada como um grafo dirigido em termos das sequencias ordenadas das suas atividades [1]. Tomando uma atividade como um vértice, uma sequência de vértices é um caminho de um grafo. A mineração de padrões em caminhos podem revelar tanto estratégias comuns para a execução da tarefa quanto comportamentos anômalos que indicam

[1]mscfaria@yahoo.com
[2]klaus.de.geus@gmail.com
[3]sergioscheer@gmail.com

um erro do aprendiz . O objetivo deste trabalho é avaliar a eficiência de diferentes métodos de agrupamento de grafos aplicados aos registros de interação de usuários de um sistema virtual de treinamento profissional (RV2) segundo níveis de execução da atividade "substituição de isolador de pedestal"em subestações elétricas.

1.1 Grafos

Um grafo dirigido G pode ser representado algebricamente por sua matriz de adjacência A_{nXn} cujas componentes A_{ij} indicam uma conexão $i \sim j$ entre dois vértices do grafo G (Figura 1). A partir da matriz de adjacência define-se a matriz laplaciana $L = A - D$ e laplaciana normalizada $L = D^{-\frac{1}{2}}LD^{-\frac{1}{2}}$. Em um grafo com arestas não ponderadas, D é a matriz diagonal dos graus dos vértices com $D_{ii} = \sum_{j=1}^{n} A_{i,j}$ igual ao número de arestas que incidem no vértice V_i. $P = \{V_1^G, V_2^G, ..., V_k^G\}$ é o caminho simples de comprimento k sobre G dado por uma sequência de vértices e as arestas que os conectam tal que não exista repetição de vértices.

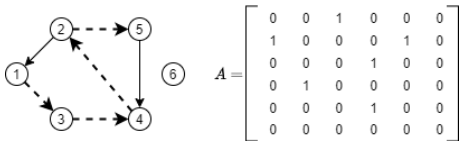

Figura 1: Grafo dirigido G e sua matriz de adjacência

De modo geral, o grau de similaridade Sim e a distância d entre dois grafos,G e G', estão relacionados pela expressão $Sim(G, G') = 1/(1 + d(G, G'))$, onde quanto menor for a distância $d(G, G')$, maior será o grau de similaridade entre G e G', e $Sim(G, G') \in [0, 1]$. As definições de distâncias são sensíveis em capturar características locais, como a detecção de anomalias entre as conexões dos vértices, ou características globais como a identificação de comunidades entre os vértices do grafo [8].

1.2 Distâncias

A distâncias apresentadas a seguir estão disponíveis na biblioteca *NetComp* para *Python* [7] .

Distâncias	*Matrizes e vetores*
$d_\lambda(G,G') = \left(\sum_{i=1}^{k}(\lambda_i^M - \lambda_i^{M'})^p\right)^{\frac{1}{p}}$	A , $L = A-D$, $L = D^{-\frac{1}{2}}LD^{-\frac{1}{2}}$
$d_{GED}(G,G') = \|A - A'\| = \sum_{i,j}^{n} \|A_{i,j} - A'_{i,j}\|$	A
$d_{VEO}(G,G') = 2\frac{\|V_G \cap V_{G'}\| + \|E_G \cap V_{G'}\|}{\|V_G\| + \|V_{G'}\| + \|E_G\| + \|E_{G'}\|}$	$\|V\|$ *e* $\|E\|$
$d_{Deltacon}(G,G') = \left(\sum\sum(\sqrt{S_{ij}} + \sqrt{S'_{ij}})\right)^{\frac{1}{2}}$	$S = [I + \epsilon D_1 - \epsilon A_1]^{-1}$
$d_{res}(G,G') = \left(\sum_{i=1}^{k}(R - R')^p\right)^{\frac{1}{p}}$	$R = diag(L^\dagger)1^T + 1diag(L^\dagger)^T + 2L^\dagger$
$d_{Netsim}(G,G') = \sum \frac{\|s - s'\|}{s + s'}$	vetor s: *"signature"*

Figura 2: Distâncias entre grafos da biblioteca *NetComp*

DOI: 10.5540/03.2021.008.01.0351

Distâncias espectrais (distâncias λ) [7] O espetro de uma matriz M é a sequencia ordenada de seus autovalores λ_i^M. As distâncias espectrais são invariantes a permutação dos rótulos dos vértices e a escolha de valores de k na sequência de autovalores de M reflete a estrutura global do grafo.

Distância de edição de grafos A distância de edição de grafos [7], *GED - Gaph Edit Distance*, é calculada a partir da quantidade de operações necessárias para transformar G em G' com custo mínimo de apagar ou incluir um vértice ou uma aresta. A diferença entre as matrizes de adjacência A e A' captura as edições realizadas na transformação de G em G'.

Sobreposição vértice/aresta Na sobreposição vértice/aresta (VEO - *Vertx/Edge Overlap*) [7] a similaridade S_{VEO} é calculada a partir da razão entre quantidade de vértices e arestas compartilhadas entre os grafos e a respectiva soma de vértices e arestas dos grafos.

Deltacon Deltacon [3] é um algoritmo que compara as afinidades entre os vértices dos grafos e foi desenvolvida a partir do método *Fast Belief Propagation* que modela a difusão da informação em um grafo. A distância $d_{Deltacon}(G, G')$ é calculada pela matrizes de afinidades entre os vértices de G e G'.

Resistência normalizada A resistência [2] de um grafo é calculada com base na analogia com um circuito elétrico onde as arestas representam resistores. A distância resistência normalizada entre dois grafos G e G' é calculada a partir da matriz de resistência R definida em termos da inversa generalizada da matriz laplaciana $L^\dagger$.

Netsimile Netsimile [3] calcula a similaridade entre dois grafos G e G' a partir de um vetor s das medidas estatísticas como a média, a mediana, o desvio-padrão, assimetria e curtose é a "assinatura"do grafo.

Os métodos de cálculo de similaridade diferem quanto a sua eficiência em determinar se dois grafos são similares principalmente porque as distâncias utilizadas são sensíveis a variações locais (diferença entre os vértices e vizinhança) ou globais (diferenças na rede de conexões dos vértices). A análise do comportamento de diferentes métodos fornece parâmetros de comparação com novos métodos que incorporem uma interpretação de resultados semanticamente adaptada ao contexto de aplicação. A seguir são apresentados a seção **Metodologia** e a **Análise dos resultados**.

2 Metodologia

A abordagem apresentada neste trabalho pode ser dividida em três etapas: o pré-processamento dos dados, processamento e validação dos agrupamentos (Figura 3). A ideia é comparar alguns métodos de similaridade aplicados a caminhos sobre um grafo por meio da qualidade dos agrupamentos formados.

O primeiro passo é codificar os dados de registro de interação de usuários durante a execução de uma atividade em um sistema virtual de treinamento na forma de caminhos sobre um grafo. Esta etapa envolve a definição 1^{o}) de um grafo dirigido que incorpora as regras de ordenação de execução das tarefas da atividade (grafo tarefa) e 2^{o}) dos caminhos sobre este grafo que representam a ordem coma qual cada usuário executou as tarefas da atividade. Para cada método foram calculadas as similaridades entre os caminhos dos usuários e todos os caminhos do grafo atividade. Antes de detalhar as etapas de agrupamento é apresentado o contexto de aplicação da análise proposta neste trabalho.

DOI: 10.5540/03.2021.008.01.0351

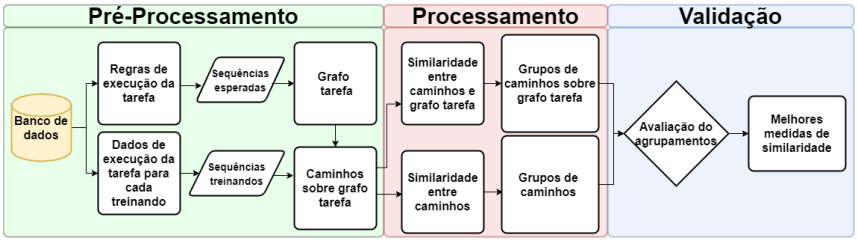

Figura 3: Etapas da Análise de Agrupamento

Contexto de aplicação do trabalho O sistema virtual de treinamento RV2 é um ambiente imersivo que reproduz uma subestação elétrica. O RV2 foi desenvolvido para treinamento a tarefa "troca de isolador de pedestal"composta por 20 atividades. Os dados de interação utilizados neste trabalho se referem ao registro de 22 sessões de treinamento e contém informações sobre o tipo e ordem na qual as atividades foram realizadas.

Pré-processamento

A modelagem dos dados foi realizada primeiro com o objetivo de mapear a tarefa em termos da ordem esperada para a execução das atividades e criar o grafo tarefa e, a segundo, para incorporar a sequência de atividades executadas por cada usuário como um caminho do grafo tarefa. [4]

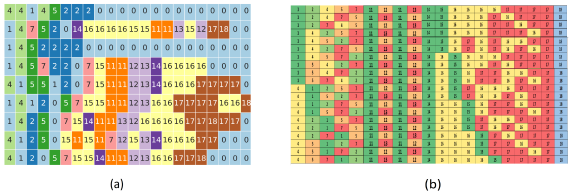

Figura 4: (a) Exemplos de sequências executadas pelos usuários. (b) Exemplos de sequências esperadas

A tarefa foi decomposta em termos de pré-requisitos da execução das atividades na forma de restrições na ordem da atividade na sequência (Figura 4 (a)). A partir destas restrições foi definido o grafo onde os vértices representam as atividades e as arestas assinalam caminhos sobre o grafo (Figura 5 (b)). Os caminhos permitidos no grafo tarefa computam 1680 sequências esperadas (Figura 4 (b)) para a execução das atividades pelos usuários do sistema constituindo portanto o gabarito da tarefa.

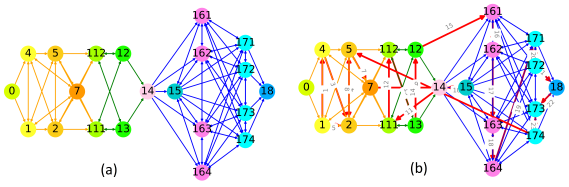

Figura 5: (a) Caminhos permitidos (b) Grafo atividade. (c) Caminho sobre grafo tarefa.

[4] A criação e manipulação dos grafos foram implementadas por meio das bibliotecas *Networkx* e os cálculos de similaridade foram feitos a partir da biblioteca *Netcomp*, ambas para *Python*.

Para efeito de comparação os caminhos foram incorporados a um grafo tarefa (Figura 5 (b)), onde os números sobre as arestas do caminho indicam a ordem em que as atividades foram executadas.

Análise de Agrupamentos A análise de agrupamento tem como objetivo identificar e agrupar os elementos de um conjunto de dados em classes, de acordo com padrões extraídos dos próprios dados, sem um conhecimento prévio dos rótulos de classe de cada elemento. Trata-se, portanto, de um método não-supervisionando de mineração de dados. A qualidade do agrupamento é mensurada pelo coeficiente de silhueta que relaciona a coesão, Q_{coe}, medida de densidade do grupo, e a separação Q_{sep}, dada pela distância entre grupos diferentes. Valores próximos de 1 indicam um agrupamento ótimo e, -1, o caso contrário.

Neste trabalho o processamento e validação dos agrupamentos, incluindo a aplicação do método *k-means*, foi realizado a partir de sua implementação na biblioteca *Scikit-learn* para *Python*. O *k-means* é inicializado com um número aleatório de grupos definidos como centroides. Em cada iteração, a posição dos centroides são atualizadas e cada elemento é atribuído a um determinado grupo segundo a menor distância em relação aos centroides. O processo tem fim quando as posições dos centroides apresentam variações abaixo de um limite estabelecido entre as iterações .

3 Resultados e discussão

A análise dos agrupamentos foi realizada comparando os caminhos entre si e os caminhos de atividades com o grafo tarefa. Para cada distância foram computadas as similaridades entre cada um dos caminhos que representam a ordem de execução das atividades pelos treinandos com todos os caminhos esperados sobre o grafo tarefa. A figura 6 apresenta os gráficos de similaridades entre os caminhos e o grafo para quatro distâncias. Embora o padrão de similaridades de um mesmo caminho sejam diferentes para cada distância, nota-se que a similaridade entre caminhos diferentes se mantém entre as distâncias. Por exemplo, dados os caminhos 4, 17 e7, os padrões dos caminhos 4 e 17 são mais similares do que os padrões entre os caminhos 17 e 7, ou, 17 e 7.

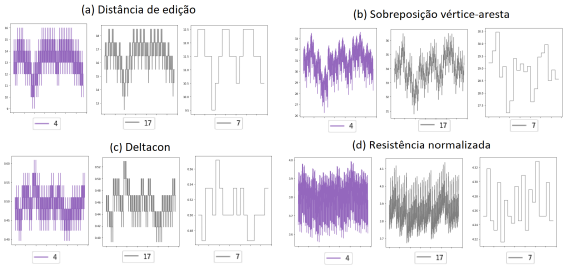

Figura 6: Similaridade entre Caminhos de atividades X Grafo Tarefa.

Os métodos são analisados separadamente e combinados, segundo sua natureza: espectrais, matriciais ou vetoriais, e a partir de duas perspectivas. Na primeira, são calculadas as similaridades entre cada caminho do aprendiz com os caminhos permitidos, ou seja, para cada execução de um aprendiz é calculada a similaridade com a execução esperada; na segunda, são calculadas as similaridades apenas entre os caminhos dos aprendizes. Os resultados encontrados são relacionados segundo seu coeficiente de silhueta e a variabilidade do número de caminhos em cada grupo.

Em todas as análises o número de grupos por agrupamento foi de 3, conforme estimativa calculada pelo "método do joelho". Os agrupamentos definidos a partir dos métodos matriciais

DOI: 10.5540/03.2021.008.01.0351

que comparam os caminhos e o grafo apresentaram maiores coeficientes de silhueta acima de 0,6 (Figura 7). Na comparação entre caminhos apenas o método sobreposição vértice-aresta alcançou um valor de silhueta igual a 0,66.

Caminhos	Delta CG	Res. N. CG	Sob. VA CC	Dis. edi CG	Sob. VA CG	Netsim CG	AdC CG	Lap CG	Netsim CC	AdC CC	Matr (S) CG	Lap CC	Matr CG	Res. N. CC	Esp CG
0	1	1	0	1	0	0	1	2	0	1	2	1	2	0	2
1	1	0	1	0	1	1	1	0	2	2	0	0	0	1	0
2	1	0	1	0	1	1	2	0	2	0	0	0	0	1	0
3	1	0	2	0	1	1	0	0	1	2	0	0	0	1	0
4	1	1	0	0	0	0	2	2	1	2	2	0	2	0	2
5	0	1	0	0	2	0	1	2	1	1	1	1	1	0	2
6	1	1	0	0	0	1	2	2	0	2	2	0	2	0	2
7	1	0	1	0	1	1	1	0	2	0	0	2	0	1	0
8	1	1	0	2	2	0	0	1	1	2	2	0	2	2	1
9	1	0	2	0	1	1	2	0	1	0	0	0	0	1	0
10	0	1	0	0	0	0	1	2	1	1	1	1	1	0	2
11	2	2	0	0	0	2	0	1	1	2	1	0	1	2	1
12	0	2	0	2	0	2	2	1	1	1	1	0	1	2	1
13	1	2	0	2	2	0	2	1	0	1	1	1	1	2	1
14	0	2	0	0	2	0	0	2	1	2	1	0	1	2	1
15	1	1	0	2	0	0	1	2	1	1	2	1	2	0	2
16	0	2	0	2	0	0	1	1	0	1	1	2	1	2	1
17	1	1	0	1	0	0	2	2	1	1	2	2	2	0	2
18	1	1	0	2	0	0	1	2	0	1	2	1	2	2	2
19	1	1	0	2	2	0	1	2	0	1	2	1	2	2	2
20	0	1	0	0	2	0	1	1	0	1	1	2	1	2	1
21	0	2	0	2	2	2	2	1	1	2	1	0	1	2	1
Silhoutte	0,69567029	0,66688622	0,64831591	0,63203793	0,60736766	0,56942606	0,54841284	0,51906984	0,47112764	0,46092497	0,42212517	0,41832509	0,38541204	0,37393503	0,36711146
PCA	1	1	2	1	1	1	1	2	2	2	1	3	1	3	2

Figura 7: Agrupamentos e coeficientes de silhueta

A maior variabilidade do número de caminhos nos grupos acompanhou um maior valor do coeficiente de silhueta. Exceção a regra são os métodos resistência normalizada e sobreposição vértice-aresta que apresentam alto valor de silhueta e baixa variância no número de caminhos em grupos. A visualização dos grupos em duas dimensões foram geradas a partir das suas componentes principais (PCA).

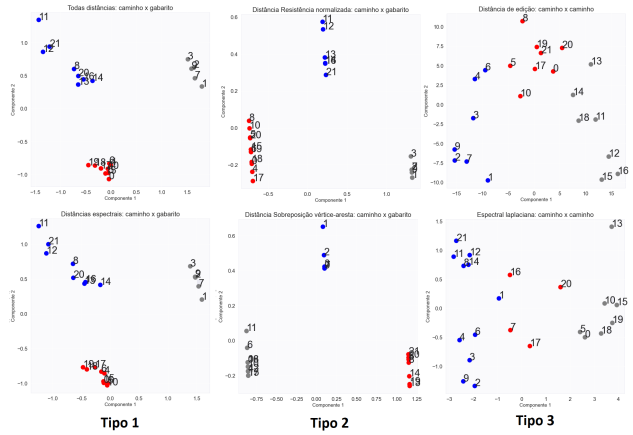

Figura 8: Agrupamentos entre caminhos

Conforme os gráficos da figura 8, três diferentes tipos de agrupamento foram obtidos. O primeiro, formado a partir de métodos aplicados entre caminhos e grafo, com coeficiente de silhueta acima de 0,6. Os grupos formados são relativamente densos e bem separados, com coeficientes de silhueta acima ou próximos da média do agrupamento; o segundo tipo de agrupamento tem representantes com diferentes valores de silhueta médio, variando de 0,3 a 0,64, mas com o coeficiente de silhueta dos grupos próximos ou abaixo da média do agrupamento. No terceiro tipo de agrupamento, caracterizado por valores de silhueta variando entre 0,4 e 0,2, indica que estes métodos não agruparam adequadamente os caminhos. Para estes agrupamentos os valores de silhuetas para os grupos ficaram abaixo da média do agrupamento. Em relação a todos os métodos, ou suas combinações, os agrupamentos que apresentaram melhores resultados foram aqueles obtidos

DOI: 10.5540/03.2021.008.01.0351

pelos métodos resistência normalizada e sobreposição vértice-aresta aplicados entre os caminhos e o grafo.

4 Conclusões

Neste trabalho foi realizada a análise de diferentes métodos para o cálculo de similaridade, e respectivas distâncias, aplicada ao agrupamento de dados de interação de usuários de um sistema virtual de treinamento. Embora este estudo tenha identificado as melhores medidas de similaridades para avaliação do conjunto de dados, os grupos formados são de difícil interpretação no contexto da aplicação pois as distâncias utilizadas foram originalmente propostas para a comparação topológica de grafos com base na conectividade dos seus vértices. No caso em que um grafo representa uma tarefa instrucional a diferença entre um caminho e o grafo de caminhos esperados não é computada segundo o erro na execução da tarefa. Esta ressalva coloca um problema de adequação semântica da similaridade calculada. Uma alternativa para futuras investigações seria criar uma nova medida de similaridade cuja distância capture o erro, como um desvio de um caminho esperado, e seu respectivo custo de correção.

Agradecimentos

Este trabalho foi desenvolvido junto ao Grupo de Investigação OneReal no âmbito do projeto de P&D PD-6491-0299/2013 proposto pela Copel Geração e Transmissão S.A., sob os auspícios do Programa de P&D da Agência Nacional de Energia Elétrica (ANEEL).

Referências

[1] Hao, J., Shu, Z., von Davier, A. Analyzing Process Data from Game/Scenario-Based Tasks: An Edit Distance Approach. Journal of Educational Data Mining, 7(1), 33-50, 2015. https://doi.org/10.5281/zenodo.3554705

[2] Klein, D.J. and Randić, M. Resistance distance. J Math Chem 12, 81–95, 1993. https://doi.org/10.1007/BF01164627

[3] Koutra, D., Faloutsos, C., Han, J., Getoor, L., Wang, W. and Gehrke J. Individual and Collective Graph Mining: Principles, Algorithms, and Applications , Morgan Claypool, 2017. https://doi.org/10.2200/S00796ED1V01Y201708DMK014

[4] Sottilare, R., Graesser, A., Hu, X., Olney, A., Nye, B. and Sinatra, A. Design Recommendations for Intelligent Tutoring Systems: Volume 4-Domain Modeling., 2016

[5] Tan, P., Steinbach, M. and Kumar, V. Introdução ao Data Mining: mineração de dados. Ciência Moderna, Rio de Janeiro, 2009. 900 p. ISBN 9788573937619.

[6] Tantardini, M., Ieva, F., Tajoli, L. et al. Comparing methods for comparing networks. Sci Rep 9, 17557 (2019). https://doi.org/10.1038/s41598-019-53708-y

[7] Wills, P. and Meyer, F.G. Metrics for graph comparison: A practitioner's guide. PLoS ONE 15(2): e0228728. 2020. doi:10.1371/journal.pone.0228728

[8] Zager, L. A. and Verghese, J. C. Graph similarity scoring and matching, Applied Mathematics Letters, Volume 21, Issue 1, Pages 86-94, , 2008. ISSN 0893-9659, https://doi.org/10.1016/j.aml.2007.01.006.

