Estudo de métricas de *ranking* na rede de citação de artigos científicos Cora

Samuel O. S. Bianch[1], **Juliana S. Silva**[1], **João G. R. Silva**[3]

[1]Instituto Federal de Educação, Ciência e Tecnologia de Mato Grosso (IFMT)
Rua Zulmira Canavarros, 95, Centro, Cuiabá - MT - Brasil

[2]Instituto Federal de Mato Grosso - Campus Pontes e Lacerda (IFMT)
Caixa Postal 99 – 78.250-000 – Pontes e Lacerda – MT – Brasil

samuelbianch38@gmail.com, joaogabriel.comp@gmail.com

Resumo. *A publicação de artigos científicos é uma ação essencial no meio acadêmico. Contudo, definir qual artigo ou publicação científica tem a maior relevância entre os autores é um desafio para os pesquisadores. Assim, o objetivo deste estudo é analisar a aplicação de algumas métricas de ranking em uma base de dados de citações, denominada Cora, na perspectiva de verificar, qual a métrica apresenta rankings mais próximos aos rankings gerados por outras métricas. A métrica que mais se correlaciona às demais tende a ser uma métrica que, de um modo amplo, se aproxima da classificação de outras considerações matemáticas utilizadas por outras métricas, o que a torna uma boa indicação quando necessário a classificação de postos para bases de dados de artigos científicos e suas citações. Para as comparações entre as métricas foram utilizados coeficientes de correlação. Resultados quantitativos indicaram que as métricas Betweenness e Grau são as que mais se correlacionam às demais, o que as tornam boas indicações para o problema.*

Abstract. *The publication of scientific papers is an essential action in academia. However, establishing which of the papers is the most relevant between the authors is a researcher's challenge. Therefore, the objective of this study is to analyze the application of some ranking metrics in a citation database, called Cora, to verify between a collection of ranking metrics, with different considerations to the calculations of the ranks, which metric demonstrates a list of rankings closer to a list generated through other metrics. This metric tends to be a metric that broadly speaking approximates itself to the ranking of other mathematical considerations used by other metrics, which makes it a good option when the classification of ranks is needed for databases of scientific articles and citations.To implement the comparison between the metrics, It was utilized correlation coefficients. Quantitative outcomes indicated that the metrics betweenness and Grau were the ones that correlate the most to other metrics, making these two good indications for the issue stated.*

1. Introdução

As citações em artigos científicos partem de diversas interações entre autores, sendo uma parte fundamental no campo da ciência. A citação torna uma afirmação mais

confiável, transformando-se em uma declaração mais forte do que quando não há uma fonte [Zhang et al. 2018].

Uma rede de citação é uma relação de apontamento para um conceito ou trabalho já existente no campo científico [Calazans et al. 2015]. Nessas redes encontram-se cientistas que, em algum momento, utilizaram conceitos já abordados por outros colegas, sendo uma ferramenta essencial na ciência, concernindo o crédito ao autor e o reaproveitamento de teorias para permitir a evolução da ciência [Liu et al. 2019].

Em meio a um número extenso de citações é possível identificar grupos de pesquisadores que possuem ligações de citações em artigos científicos. Assim, este grupo de interações pode ser definido como uma rede social, uma vez que a mesma pode ser representada como uma ligação entre pessoas [Silva et al. 2020]. Diante dessa realidade, as citações encontradas em artigos podem ser representadas por um grafo, por meio da modelagem computacional, que consiste em uma abstração da realidade. Para que se compreenda como os cientistas se relacionam entre si, por intermédio das Redes Complexas, torna-se possível aplicar um estudo ranqueando os artigos por métricas de *ranking* de vértices, existentes na literatura [Silva 2014].

Cora é uma base de dados de artigos científicos, que foi modelada como uma rede de citação, sendo um exemplo de como os pesquisadores se relacionaram em seus trabalhos, com outros cientistas [Kunegis 2017]. Nessa rede, os artigos são representados por vértices e as arestas são mapeadas pelas citações entre os artigos. Dessa forma, pode-se estabelecer relações de cunho científico entre os autores. Vale ressaltar que a rede de citação Cora possui arestas apontando apenas para artigos da própria rede, excluindo citações fora do escopo.

Avaliando a rede Cora e na busca por uma forma de se obter o artigo mais influente da rede, propõe-se, neste trabalho, a aplicação de diferentes métricas de *ranking* à essa base e a comparação dos *ranking* gerados por cada uma dessas métricas, no qual o objetivo da pesquisa é verificar qual métrica mais se correlaciona com as demais, consequentemente, a métrica que abarca uma forma mais eficiente de classificação de artigos dado que cada métrica possui considerações diferentes para suas classificações.

2. Aspectos Teóricos

2.1. Redes Complexas e Redes de Citação

Uma Rede Complexa é modelada por meio de um grafo [Metz et al. 2007], que consiste em uma representação gráfica de interações entre indivíduos, máquinas, cidades, aeroportos e outras demais aplicações. Neste estudo, foi analisada uma rede de citação de artigos científicos - Cora. Assim, para a modelagem da rede foi utilizado um grafo direcionado, ou seja, existe uma direção na relação.

Uma rede de citação é um modelo em que representa citações entre autores de artigos científicos como disposto na Figura 1. Esta rede é modelada como uma lista de artigos (nós) relacionados entre si, por meio de citações (arestas). Assim, o nó x (Artigo 1) possui um link para o nó y (Artigo 2) e assim sucessivamente.

Além desses conceitos, é importante descrever o conceito de assortatividade, que é a tendência de que um vértice tem a se ligar com outros vértices que possuem um perfil similar ou diferente, considerando uma característica comum entre eles

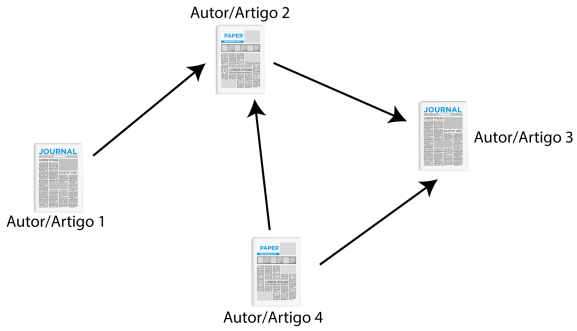


Figura 1. Representação de uma rede de citação

[Barbosa et al. 2014]. Nesta rede, uma alta assortatividade significa que os autores de mesmo nível se citam.

2.2. Métricas de Ranking

As métricas de *ranking* são medidas vastamente utilizadas para organizar os vértices em postos de acordo com o que é definido como importante para a análise, sendo bastante consumida no campo das Ciência de Dados. Neste estudo foram abordadas quatro métricas de centralidade: Grau (*ranking*), centralidade de proximidade (*Closeness*), centralidade de intermediação (c) e a centralidade de autovetor (*Eigenvector*). Para enriquecimento do estudo, no experimento também é utilizado dois algoritmos de análise de *links Hyperlink-Induced Topic Search* (análise de links), que a partir de agora será tratado como *HITS algorithm*, que são as métricas: *Hub* e *Authority*. Como também, o ranqueamento da página (*PageRank*). Cada uma das métricas utilizadas neste estudo são apresentadas a seguir.

O Grau é uma medida de centralidade simples que define o nó mais importante como aquele que possui mais arestas apontando ou saindo deste nó. Pode ser calculado a partir do número de arestas que apontam para este nó, como também as arestas que partem deste vértice.

A centralidade de *Closeness*, classifica o nó de acordo com a menor distância em relação aos outros nós da rede. Esses nós possuem o menor caminho de um ponto x a um ponto y da rede [Stephenson and Zelen 1989].

A centralidade de *Betweenness*, classifica os vértices de acordo com o seu grau de intermediação na rede, ou seja, o cientista que mais faz pontes na rede terá uma melhor colocação [Machado and Boeres 2016]. A centralidade de *Eigenvector*, é difundida principalmente em estudos algébricos e adaptada as redes complexas. Partindo de um conceito de matriz simétrica, essa medida faz a relação entre os nós adjacentes de acordo com as pontuações dentre as conexões [Bonacich 1987].

Os algoritmos de concentração e autoridade (*Hub e Authority*) consistem em duas métricas que se interligam entre si, no qual um nó com uma pontuação alta de concentração é um nó que possui muitas arestas, porém quanto maior for a autoridade dessas arestas, maior será sua colocação. A autoridade, que por sua vez, é a reputação de cada vértice, para tanto, quanto mais munidos de concentrações de nós apontando para este nó, melhor será sua colocação [Kleinberg et al. 1998].

O ranqueamento da página (*PageRank*) é uma medida de centralidade de páginas da internet em que seu algoritmo classifica os nós de acordo com a qualidade em que os outros nós apontam para ele [Page et al. 1999]. Entretanto, este algoritmo foi adaptado para as redes de citação. Para tanto, considerando artigos como páginas e citações como o *link* entre essas páginas. Considerações matemáticas e conceituais de cada métrica podem ser encontradas em cada literatura citada.

2.3. Coeficientes de Correlação

Os coeficientes de Correlação são utilizados para definir o quanto uma classificação é próxima da outra. Neste trabalho, foram utilizados dois coeficientes de postos: *Spearman* e *Kendall-Tau*, estes serão descritos abaixo.

O coeficiente de *Spearman*, popularmente conhecida na matemática como uma derivação do coeficiente de *ranking*, usa o ρ como um retorno de quanto um ranking é próximo ao outro [Spearman 1910]. É um comparador eficiente pois utiliza o peso das diferenças de posições, sendo um dos mais indicados para a esta pesquisa. Seu espaço amostral está entre [-1, 1], quanto mais próximo de (-1), mais divergente é a correlação, já para (1) acontece o oposto, mais próximo é um *ranking* de outro.

O coeficiente *Kendal-Tau*, que basicamente retorna a diferença entre pares concordantes modularizada [Ghent 1963], tem o mesmo espaço amostral que o coeficiente anteriormente citado, facilitando assim a compreensão e visualização gráfica dentre as disparidades de classificações. As considerações matemáticas dos coeficientes de correlação utilizados podem ser encontrados nos artigos supracitados nos parágrafos acima.

3. Material e Métodos

Visando alcançar o objetivo do trabalho, foi realizado um estudo de caso aplicado na rede de citação Cora, com o intuito de verificar a aplicação de algumas métricas de *ranking* sobre a base de dados disponível, visando encontrar uma métrica que classifique a importância de um artigo científico, baseado na rede, de modo mais próximo às demais. Para tanto, o trabalho foi dividido em três etapas, a saber: armazenamento da rede de dados Cora, aplicação e escolha das métricas de *ranking* à rede, comparação por meio, de coeficientes de correlação, dos *rankings* obtidos por cada uma das métricas selecionadas.

A primeira etapa da pesquisa foi realizada mediante um armazenamento de dados retirados da plataforma do grupo de estudos de pesquisadores da University of Koblenz–Landaus na Alemanha, o KONECT Project [Kunegis 2013].

De modo subsequente, uma análise sobre características da rede analisada foi realizada, utilizando propriedades das Redes Complexas. Em seguida, para a segunda etapa, foram selecionadas as métricas de *ranking* a serem aplicadas à rede de artigos Cora. Nesse sentido, as métricas selecionadas foram: Grau, *Closeness, Betweenness, Eigenvector, Hub, Authority, PagRank*. A escolha dessas métricas baseia-se no fato de que cada uma realiza considerações matemáticas distintas em seus cálculos, além de existir experimentos na literatura que promoveram tais comparações [Silva et al. 2015].

Na etapa de aplicação dessas métricas foi utilizada a linguagem de programação *Python* com o auxílio da biblioteca *Igraph* e um código base, que foi manipulado para este trabalho, disponibilizado pelo *GitHub* de [Silva et al. 2015].

Para a comparação dos postos foram utilizados os coeficientes de correlação de postos de [Ghent 1963] e [Spearman 1910]. A escolha por esses coeficientes de correlação baseou-se no fato de serem tradicionais da área de estatística [Myers and Sirois 2004], [Bolboaca and Jäntschi 2006] e também utilizados atualmente por [Flight et al. 2022].

A confrontação das métricas de *ranking* utilizadas foi realizada de forma a comparar os *rankings* gerados por cada uma das métricas, par a par. Ou seja, o *ranking* gerado por cada métrica foi comparado com os *rankings* de todas as outras métricas. Para tanto, cada comparação foi realizada de dois modos: utilizando o coeficiente de correlação de *Pearson* e o coeficiente de comparação de *Kendall-Tau*. Os resultados dessas comparações foram dispostos na seção de Resultados e Discussão com o auxílio de gráficos de barra e de calor.

4. Resultados e Discussão

Diante dos métodos encontrados na literatura para abarcar um estudo na área de Redes Complexas, buscou-se definir uma métrica de *ranking* de vértices que possuísse mais elementos matemáticos do que as demais métricas, para classificar os artigos científicos, de acordo com a sua importância para a rede. Além disso, buscou-se identificar, conforme o cálculo da métrica mais similar, qual era o critério que mais se mostrava valioso para um *paper* ser melhor ranqueado, segundo a metodologia, propõe-se nesta seção a apresentação dos resultados. Dessa forma, chegou-se aos resultados descritos a seguir.

Os dados disponíveis na base de dados de artigos científicos da rede de citação Cora, como enunciado na Tabela 1, possuem 91.500 artigos (vértices) e 23.167 citações (arestas). Em uma Rede Complexa direcionada é muito comum a existência de mais arestas do que vértices. Entretanto, quando observa-se a rede de citação Cora, é constatado esse desequilíbrio, o que significa que esta rede possui mais artigos publicados, do que artigos citados na própria rede.

Propriedade	**Valor**
Número de vértices/artigos	91500
Número de arestas/citações	23167
Assortatividade	0.0047
Transitividade Média	0.3071

Tabela 1. Dados rede Cora

Diante da Tabela 1, ao analisar a propriedade da assortatividade, as características da rede apresentam índices baixos. Infere-se que, devido a autores menos reconhecidos citarem cientistas mais conhecidos, vértices de diferentes níveis (importância ou número de citações), se citem mais frequentemente. De face à transitividade média, são encontrados pequenos grupos de autores, o que faz com que a tendência da rede, em criar comunidades, seja baixa para média. Cabe ressaltar, que as duas propriedades (assortatividade e transitividade média) são dispostas de valores entre 0 e 1, no qual 0 representa o valor mínimo da propriedade e 1 apresenta o valor máximo.

Ranqueando os vértices, de acordo com a relevância de cada artigo, pode-se analisar qual publicação possui a maior importância dentre todas as outras. Para tanto, a

Figura 2 apresenta um gráfico de calor, no qual as métricas são comparadas, uma a uma, utilizando coeficientes de correlação. Nessa Figura, valores de correlação próximos a 1 (maior correlação) são apresentados em preto e valores próximos a -1 (menor correlação) em branco. Nota-se, na diagonal principal, os valores máximos de correlação em preto (1), pois o *ranking* gerado pela métrica é comparado com ele próprio. Vale reforçar que, a matriz triangular inferior apresenta os resultados das correlações, utilizando o coeficiente de *Spearman*, e, na diagonal superior, o coeficiente *Kendall-Tau*.

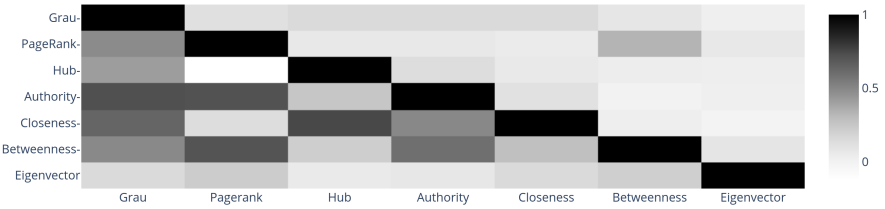
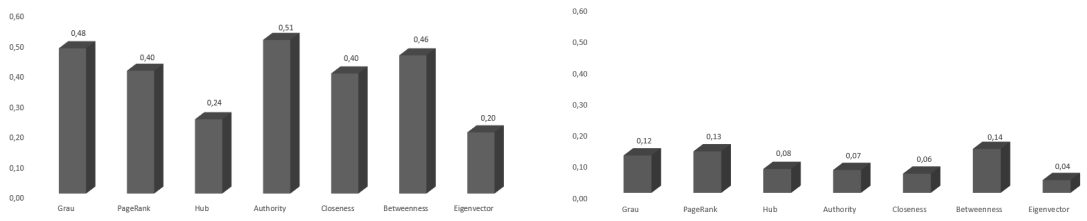


Figura 2. Comparativo das métricas de *ranking*

Como esboçado na Figura 2, é encontrado uma maior correlação nos *rankings* comparados utilizando o coeficiente de *Spearman*, isto fica evidenciado por conta do seu cálculo usufruir do peso das diferenças de colocações entre os postos. Ademais, o coeficiente de *Kendall-Tau* se mostra muito sensível a estas diferenças de colocações entre os artigos científicos.

Outro resultado da análise, disposto na Figura 3, refere-se à média de correlação de cada métrica, comparada às demais, utilizando os dois coeficientes de postos: (a) *Spearman* e (b) *Kendall-Tau*. Observa-se, na Figura 3(a) que a métrica *Authority*, Grau e *Betweenness* são as métricas, utilizando o coeficiente de correlação de *Spearman*, que mais se correlacionam às demais. Logo, a classificação dos artigos, por estas métricas, pode unir considerações de outras métricas de *ranking*. Com relação à Figura 3(b), as métricas que mais se correlacionam com as demais são: *PageRank*, *Betweenness* e Grau, comparadas utilizando o coeficiente de *Kendall-Tau*.



(a) *Spearman* (b) *Kendall-Tau*

Figura 3. Média das métricas correlacionadas

Considerando a análise apresentada nas Figuras 3 e 4, pode-se inferir que, caso seja necessário, a indicação de uma métrica para classificação de importância de um artigo

científico na rede Cora, rocomenda-se a utilização de *Betweenness*, em primeiro ponto ou simplesmente Grau, em uma rede de citação que considera basicamente o número de citações. A recomendação baseia-se no fato destas métricas ficarem entre as duas que mais se correlacionam utilizando coeficientes de correlação distintos.

Isso posto, salienta-se que o presente estudo, diverge na métrica que mais se correlaciona com as outras dos estudos modelo [Silva et al. 2020] e [Silva et al. 2015]. Esse fato confirma que as características e topologia de diferentes redes complexas influenciam em sua correlação de postos.

5. Conclusão

Apresentar qual artigo científico tem a maior relevância no meio científico não é uma tarefa trivial. Isso se deve aos diferentes parâmetros que estão em questão para definir qual será o artigo mais influente. Considerando o contexto apresentado, as métricas *Betweenness* e Grau foram as que mais se destacaram diante das demais métricas, sendo as mais recomendadas para o problema.

Este trabalho propôs um estudo de caso da base de dados de artigos científicos Cora sobre o comportamento de métricas de *ranking* visando selecionar uma métrica que mais se correlaciona às demais. Esta métrica, de modo intrínseco, tende a se parecer com o maior número de métricas com cálculos e considerações bastante distintas no âmbito matemático e de topologia da rede. Logo, sendo a mais indicada quando necessário a utilização de uma métrica para classificação de artigos.

Um limitador do trabalho é que faz-se necessário um espaço amostral de artigos maior para que se conclua, com ainda mais propriedade, qual a métrica se correlaciona melhor com as outras. Ademais, as bases de dados de artigos científicos, cujas citações são evidenciadas em formato numérico, na literatura mostram-se escassas, evidenciando a necessidade de grupos de estudos modelarem redes de citações maiores e mais amplas.

Como trabalhos futuros, pretende-se modelar uma rede de citações com autores brasileiros, como do ENCompIF, por exemplo, e utilizando artigos científicos publicados no Brasil. Também considera-se de grande valia a utilização de outras métricas de *ranking*, bem como coeficientes de correlação para realização de estudos similares com ainda mais considerações.

Referências

Barbosa, L. M., Attux, R., and Godoy, A. (2014). Uma análise de assortatividade e similaridade para. *[sn]*.

Bolboaca, S.-D. and Jäntschi, L. (2006). Pearson versus spearman, kendall's tau correlation analysis on structure-activity relationships of biologic active compounds. *Leonardo Journal of Sciences*, 5(9):179–200.

Bonacich, P. (1987). Power and centrality: A family of measures. *American Journal of Sociology*, 92(5):1170–1182.

Calazans, M. M., Furtado, R. L., and Tomaél, M. I. (2015). Redes de citação: estudo de rede de pesquisadores a partir da competência em informação. *Em Questão*, 21(2):181–202.

Flight, R. M., Bhatt, P. S., and Moseley, H. N. (2022). Information-content-informed kendall-tau correlation: Utilizing missing values. *bioRxiv*.

Ghent, A. (1963). Kendall's "tau" coefficient as an index of similarity in comparisons of plant or animal communities. *The Canadian Entomologist*, 95(06):568–575.

Kleinberg, J. M. et al. (1998). Authoritative sources in a hyperlinked environment. In *SODA*, volume 98, pages 668–677. Citeseer.

Kunegis, J. (2013). KONECT – The Koblenz Network Collection. In *Proc. Int. Conf. on World Wide Web Companion*, pages 1343–1350.

Kunegis, J. (2017). Cora network dataset – KONECT http://konect.cc/networks/subelj_cora/.

Liu, H., Kou, H., Yan, C., and Qi, L. (2019). Link prediction in paper citation network to construct paper correlation graph. *EURASIP Journal on Wireless Communications and Networking*, 2019(1):1–12.

Machado, A. M. and Boeres, M. C. S. (2016). Aplicação de medidas de centralidade e análise da estrutura da rede brasileira de financiamento de campanha eleitoral de 2014. *XLVIII SBPO SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL. XLVII., Vitória. Anais do XLVIII Simpósio Brasileiro de Pesquisa Operacional, Vitória, ES*.

Metz, J., Calvo, R., Seno, E. R., Romero, R. A. F., Liang, Z., et al. (2007). Redes complexas: conceitos e aplicações.

Myers, L. and Sirois, M. J. (2004). Spearman correlation coefficients, differences between. *Encyclopedia of statistical sciences*, 12.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Silva, J. G. R., Xavier, C. R., da Fonseca Vieira, V., and de Carvalho, I. A. (2015). Estudo comparativo de métricas de ranqueamento em redes complexas utilizando coeficientes de correlaçao. In *Congresso Brasileiro de Inteligência Computacional*.

Silva, J. S. (2014). Métricas de análise de redes sociais e sua aplicação em redes de interação biológicas: metodologia de aplicação e estudos de caso. *Tese (Doutorado em Sistemas Digitais) - Escola Politécnica, Universidade de São Paulo, São Paulo, 2014*.

Silva, S. O., Goulart, B. O., Schettini, M. J. M., Xavier, C., and Silva, J. G. (2020). Estudo comparativo de métricas de ranking em redes sociais. In *Anais do VII Encontro Nacional de Computação dos Institutos Federais*, pages 53–60. SBC.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 1904-1920*, 3(3):271–295.

Stephenson, K. and Zelen, M. (1989). Rethinking centrality: Methods and examples. *Social networks*, 11(1):1–37.

Zhang, A. X., Ranganathan, A., Metz, S. E., Appling, S., Sehat, C. M., Gilmore, N., Adams, N. B., Vincent, E., Lee, J., Robbins, M., et al. (2018). A structured response to misinformation: Defining and annotating credibility indicators in news articles. In *Companion Proceedings of the The Web Conference 2018*, pages 603–612.