Probabilidade de visitar um ponto turístico: uma análise usando grafos e PageRank

Julia Rezende Gomes Rocha
Engenharia de Computação
CEFET - MG - Campus V
Divinópolis - MG - Brasil
juliarezende34@gmail.com

***Resumo*—Este artigo apresenta a análise de pontos turísticos de cidade de Los Angeles e tem como objetivo obter a probabilidade de visitar outros lugares a partir de um ponto inicial. Para tal, foi utilizada a teoria de grafos para modelar o problema e o algoritmo PageRank para o cálculo das probabilidades. Como o uso do PageRank, as sugestões para a próxima visita são baseadas equilíbrio entre o número de visitantes da origem e do destino, e da distância entre eles.**

***Abstract*- This article presents the analysis of tourist spots in the city of Los Angeles and aims to obtain the probability of visiting other places from an initial point. For this purpose, graph theory was employed to model the problem, and the PageRank algorithm was utilized for probability calculations. With the use of PageRank, suggestions for the next visit are based on a balance between the number of visitors at the origin and destination, as well as the distance between them.**

***Index Terms*—Grafos, PageRank, Turismo, Python, Networkx**

I. Introdução

Montar um roteiro de turismo é uma tarefa árdua, seja o interessado um indivíduo organizando sua própria viagem ou uma agência de viagens auxiliando um cliente. Organizar uma visita à um lugar desconhecido depende de muitos fatores, como tempo, distância, custo e interesse dos viajantes, por isso faz-se necessário um método facilitador para essa atividade, que possa indicar lugares a serem visitados com base em características dos pontos turísticos e suas relações.

Este trabalho tem como objetivo analisar uma massa de dados referente à pontos turísticos da cidade de Los Angeles e descobrir a probabilidade de se visitar outros a partir de um ponto inicial, além de considerar se esta é impactada pelas características dos lugares. Tem-se como hipótese que quanto maior a distância entre os pontos, menor a probabilidade de caminhamento entre eles. Com essa análise, será possível montar um roteiro de visitação personalizado, sem realizar muito esforço e de boa qualidade.

As seções deste artigo são divididas da seguinte maneira: na seção 2 acontece a contextualização do problema, de onde surgiu a necessidade de analisá-lo e uma introdução à grafos. Na seção 3, evidencia-se os trabalhos correlatos ao problema e como esses impactam esta pesquisa. A seção 4 é responsável pela descrição do método usado para analisar o problema, na seção 5 os resultados são apresentados e nas seções 6 e 7 são indicadas as considerações finais e perspectivas para trabalhos futuros, respectivamente.

II. Contextualização

O turismo é uma das atividades econômicas fundamentais para o mercado global, contribuindo significativamente para a geração de empregos, movimentação da hotelaria e do comércio local. No Brasil, por exemplo, muitas cidades respiram turismo e este é base da sua economia. Em 2019, 7,7% do Produto Interno Bruto (PIB) brasileiro foi proveniente das atividades turísticas [1]. A partir da suma importância do turismo para a economia do Brasil, é necessário investigar, aprimorar e até mesmo automatizar os procedimentos da categoria, visando rapidez em atendimentos de agências de viagens por exemplo, aumentando o número de clientes atendidos.

Os pontos turísticos e as distâncias que os conectam podem ser representados como vértices e arestas de um grafo, devido a individualidade dos pontos e um claro critério de conexão. Um grafo G = (V, E) consiste em V, um conjunto não vazio de vértices, e E, um conjunto de arestas. Cada aresta tem um ou dois vértices associados a ela, ligando-os [2]. As arestas podem ser relacionadas a valores, também denominados como pesos, que servem para ponderar a relação entre aqueles vértices interligados.

A tabela com museus da cidade de Los Angeles [3] contém seus nomes e seu número de visitantes por mês. Visa-se relacionar o número de visitantes de um ponto A, o número de visitantes de um ponto B e a distância entre A e B, criando um peso para a aresta que os conecta. A partir desse peso, deverá ser possível aplicar o algoritmo de PageRank, obtendo uma probabilidade de caminhamento entre A e B. A partir do ponto A, por exemplo, será possível obter os pontos mais prováveis de serem visitados a partir dali, observando o impacto dos números de visitantes e da distância entre os pontos sob as probabilidades.

III. Trabalhos correlatos

O trabalho de Kamienski, Damaceno e Mena-Chalco (2019) [4] aplica o algoritmo PageRank em um grafo para identificar pesquisadores de maior prestígio em termos de formação de recursos humanos limitados pelo número de gerações de acadêmicos. Cada vértice é um pesquisador e cada aresta uma relação de orientação acadêmica.

Já no artigo de Fernandes (2018) [5], o PageRank é utilizado para classificar candidados à uma vaga de emprego, de acordo com suas especialidades disponíveis no LinkedIn.

É possível perceber que grafos e o PageRank já são usados atualmente como um método confiável de organização e classificação em dados, então é válido estender sua aplicação para outros âmbitos, como recomendações turísticas.

IV. Metodologia

A. Limpeza de dados

Os dados dos museus são organizados na tabela original da seguinte forma: cada linha contém o nome e o número de visitantes em um determinado mês. Como há meses que não possuem os dados de todos os museus, foi selecionado para ser analisado neste trabalho o mês de julho de 2019. O critério de escolha usado foi apenas utilizar um mês que contivesse dados da maioria dos museus. Como não havia um mês com dados de todos os museus, os seguintes foram escolhidos para a análise: America Tropical Interpretive Center, Avila Adobe, Chinese American Museum, Gateway to Nature Center, Firehouse Museum, IAMLA, Pico House e Museum of Social Justice. Essa limpeza de dados não foi realizada via programação, mas sim por filtragem de dados da tabela usando o Microsoft Excel.

Nome	Número de visitantes 07/2019
America Tropical Interpretive Center	4491
Avila Adobe	24097
Chinese American Museum	2271
Gateway to Nature Center	999
Firehouse Museum	4432
IAMLA	1480
Museum of Social Justice	2911
Pico House	90

Tabela I
TABELA DE DADOS APÓS LIMPEZA.

B. Primeira conexão

Para construir o primeiro grafo, uma segunda tabela foi construída. Todos os pontos foram conectados entre si, formando duas colunas: ”N1”e ”N2”, como se estas representassem a origem e o destino de cada ligação. Os valores da coluna que representa o peso da conexão entre os museus foram calculados com a seguinte fórmula:
$N_1 =$ Número de visitantes do ”N1”;
$N_2 =$ Número de visitantes do ”N2”;
$d =$ Distância entre o ”N1”e o ”N2”.

$$Weight = \frac{|N_1 - N_2|}{d} \quad (1)$$

Esta foi a fórmula escolhida pois com ela será possível observar na razão entre as grandezas o impacto de cada uma. Por exemplo, quando maior a distância entre os museus, menor será o peso daquele caminho, o que poderá indicar menor chance de caminhamento.

N1	N2	Distância (m)	Weight
America Tropical Interpretive Center	Avila Adobe	1000	19,606
America Tropical Interpretive Center	Chinese American Museum	650	3,415385
America Tropical Interpretive Center	Gateway to Nature Center	700	4,988571
America Tropical Interpretive Center	Firehouse Museum	193000	0,000306
America Tropical Interpretive Center	IAMLA	460	6,545652
America Tropical Interpretive Center	Pico House	900	4,89
America Tropical Interpretive Center	Museum of Social Justice	600	2,633333
Avila Adobe	Chinese American Museum	750	29,10133
Avila Adobe	Gateway to Nature Center	800	28,8725
Avila Adobe	Firehouse Museum	193000	0,101891
Avila Adobe	IAMLA	140	161,55
Avila Adobe	Pico House	1000	24,007
Avila Adobe	Museum of Social Justice	650	32,59385
Chinese American Museum	Gateway to Nature Center	370	3,437838
Chinese American Museum	Firehouse Museum	193000	0,011197
Chinese American Museum	IAMLA	500	1,582
Chinese American Museum	Pico House	300	7,27
Chinese American Museum	Museum of Social Justice	350	1,828571
Gateway to Nature Center	Firehouse Museum	193000	0,017788
Gateway to Nature Center	IAMLA	500	0,962
Gateway to Nature Center	Pico House	260	3,496154
Gateway to Nature Center	Museum of Social Justice	400	4,78
Firehouse Museum	IAMLA	193000	0,015295
Firehouse Museum	Pico House	193000	0,022497
Firehouse Museum	Museum of Social Justice	193000	0,007881
IAMLA	Pico House	850	1,635294
IAMLA	Museum of Social Justice	550	2,601818
Pico House	Museum of Social Justice	750	3,761333

Tabela II
CONEXÕES E PESOS PARA O GRAFO INICIAL.

A Tabela II foi usada como base para o primeiro grafo ser construído. A linguagem de programação utilizada para tal foi Python, juntamente com suas bibliotecas Pandas, Networkx e Matplotlib, no ambiente de desenvolimento Visual Studio Code[1]. O modo de construção é descrito com o pseudocódigo a seguir:

Algorithm 1 Construção do primeiro grafo

```
dados ← ler_csv('grafo.csv')
for cada linha em dados do
    linha['Weight'] ← trocarVirgulaPorPonto(linha['Weight'])
    linha['Weight'] ← converterFloat(linha['Weight'])
end for
grafo ← criarGrafo()
for cada linha em dados do
    adicionarAresta(grafo, linha['N1'], linha['N2'], linha['Weight'])
end for
exibir_grafo(grafo)
```

[1]O código usado para a metodologia está disponível em: link do repositório.

C. Definição do PageRank

O PageRank foi criado pelos fundadores do Google Larry Page e Sergey Brin, em 1997, com o objetivo de durante a pesquisa de um usuário, indicar o link mais relevante considerando sua relevância e a importância da informação ali contida [6].

O algoritmo usa o conceito da Cadeia de Markov para classificar a relevância de cada link. O valor da importância do site é determinado pela qualidade dos links de transição para outra página, que também será analisada [7], esse procedimento prioriza as páginas que têm autores ou informações com credenciais mais robustas.

O cálculo da Cadeia de Markov tem como objetivo determinar a probabilidade de acesso a uma página, usando o "tempo discreto, um número tão grande a ponto de que essas probabilidades não se alterem mais"(Oliveira, Bovoloni, Leite Filho) [7]. Após os cálculos, considerando a estabilização do tempo, cada página terá o seu número de PageRank, que também representa a soma dos PageRank's dos links que a recomendam. Quanto mais recomendações, maior o PageRank e maior importância a página possui no momento da pesquisa.

No artigo "PAGE RANK: O FUNCIONAMENTO DA FERRAMENTA DE BUSCA DO GOOGLE", de Oliveira, Bovoloni e Leite Filho [7], há o detalhamento completo do cálculo da Cadeia de Markov com exemplificação.

D. Aplicação do PageRank

A aplicação do PageRank no contexto do grafo gerado pela Tabela II acontece pelo uso do algoritmo da biblioteca Networkx do Python, que é uma iteração que redistribui a pontuação de acordo com a estrutura do grafo. O seguinte cálculo é aplicado [8]:

$$PR(u) = (1-d) + d\left(\frac{PR(v_1)}{L(v_1)} + \frac{PR(v_2)}{L(v_2)} + ... + \frac{PR(v_n)}{L(v_n)}\right) \quad (2)$$

Tal que:

- $PR(u)$ é o valor do PageRank do nó u;
- d é um fator de amortecimento, sendo 0.85 no contexto do Google;
- $PR(v_i)$ são os valores de PageRank dos nós v_i que se conectam com u;
- $L(v_i)$ é o número de arestas de saída de v_i.

O PageRank é aplicado no grafo de acordo com o seguinte pseudocódigo:

As arestas no novo grafo são obtidas de um dicionário criado pela função do PageRank onde as chaves são os nós do grafo e os valores são as pontuações de PageRank associadas a esses nós.

Para minimizar a possibilidade de erros de visualização do caminho mais provável, implementou-se o seguinte pseudocódigo que analisa os nós do grafo e os pesos das arestas incidentes à ele, retornando o caminho mais provável a partir de certo nó:

Algorithm 2 Aplicar PageRank no grafo

```
pagerank ← nx.pagerank(G)        ▷ Calcula o PageRank para cada nó em G
novoGrafo ← criarGrafo()
for all (source, target) ∈ G.arestas() do
    weight ← G[source][target]['weight']
    novoWeight ← pagerank[target]
    novoGrafo.addAresta(source, target, novoWeight)
end for
```

Algorithm 3 Processamento de Nós e Arestas com Maior Peso

```
for all u em G.nodes() do
    incident_edges ← obterArestasIncidentes(graph, u)
    if incident_edges then
        arestaMaior ← maiorPeso(incident_edges)
        v ← obterNoLigado(arestaMaior)
        imprimir(u, v, arestaMaior[2])
    end if
end for
```

V. Resultados

A. Primeira conexão

O Algoritmo 1, juntamente com a Tabela II, produziram como o resultado a seguinte ligação entre os museus:

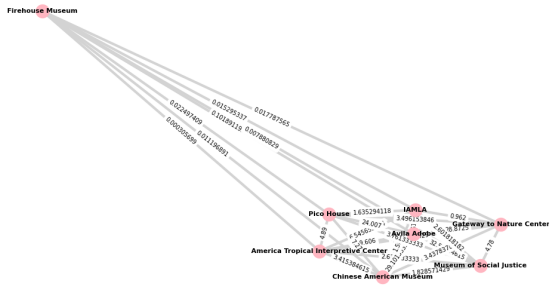

Figura 1. Grafo feito a partir das conexões descritas na Tabela II.

É possível observar um grande distanciamento do ponto "Firehouse Museum", isso acontece devido à sua grande distância em relação aos outros pontos (193000 metros). A partir da distância alta aplicada na fórmula da Equação 1, obteve-se um peso muito pequeno, distanciando esse ponto dos outros. Esse distanciamento pode ser interpretado como uma ligação fraca do "Firehouse Museum"com os outros museus, podendo indicar baixa probabilidade de caminhamento. Porém, como o PageRank não foi aplicado nesse grafo, ainda não é possível fazer afirmações.

B. Aplicação do PageRank

O Algoritmo 2, aplicado no grafo da Figura 1 nos dá o seguinte grafo direcionado, que possui o peso da aresta como as pontuações de PageRank associadas àqueles nós interligados:

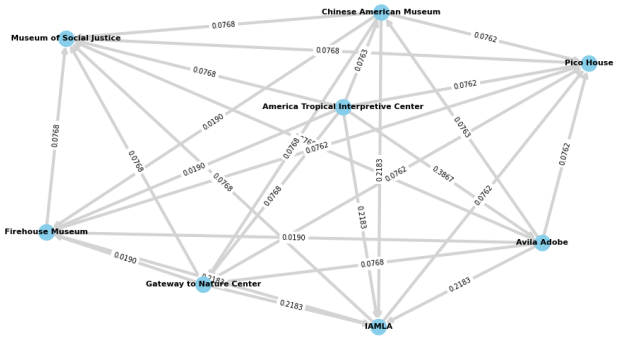

Figura 2. Grafo após aplicação do PageRank

A partir do grafo após a aplicação do PageRank, pode-se verificar visualmente o próximo ponto mais provável de ser visitado. Por exemplo, partindo do "Chinese American Museum"é mais provável que a próxima visita seja o "IAMLA", devido ao peso da aresta que os liga ser o maior entre as arestas incidentes ao "Chinese American Museum".

A partir do Algoritmo 3 foi possível obter as seguintes relações de caminho mais provável a partir de determinado nó:

Origem	Destino	Peso
America Tropical Interpretive Center	Avila Adobe	0.38671258422438365
Avila Adobe	IAMLA	0.2183059836125501
Chinese American Museum	IAMLA	0.2183059836125501
Gateway to Nature Center	IAMLA	0.2183059836125501
Firehouse Museum	IAMLA	0.2183059836125501
IAMLA	Museum of Social Justice	0.0768113326635817
Pico House	Museum of Social Justice	0.0768113326635817

Tabela III
DESTINOS MAIS PROVÁVEIS A PARTIR DE CADA UM DOS NÓS.

Analisando os destinos mais prováveis e seus dados na tabela original, é possível afirmar que a fórmula utilizada pelo algoritmo do PageRank equilibra a distância e o número de visitantes do destino, priorizando os destinos que possuem essas características balanceadas, já que estes não são sempre o mais perto ou o mais popular, mas sim nós com equilíbrio dessas características.

VI. Considerações finais

Foi possível analisar os pontos turísticos da cidade de Los Angeles com grafos e o algoritmo PageRank, descobrindo a probabilidade de visita de outros pontos a partir de um lugar inicial. Ao contrário da hipótese inicial, a maior distância entre pontos não significa que, necessariamente, a menor probabilidade de visita, pois tanto a fórmula do peso da aresta quanto o cálculo do PageRank geram um valor de equilíbrio entre os números de visitantes dos pontos e a distância entre eles.

VII. Trabalhos futuros

Para futuras expansões dessa pesquisa, considera-se adequado buscar mais detalhes dos pontos turísticos, como faixa etária da maioria dos visitantes, custo de visitação (ingressos, comida, estacionamento, por exemplo) e interesse do visitante.

Para uma aplicação mais robusta, os interesses de um usuário poderiam ser submetidos e a após análise de um conjunto maior de pontos turísticos, obter indicações de cidades ou países que contém atividades e locais que agradariam o usuário.

VIII. Referências

[1] RIBEIRO, Luiz Carlos de Santana; SANTOS, Monique Manuela Carvalho dos; SANTOS, Fernanda Rodrigues dos. Avaliação das atividades características do turismo no Brasil: 2012-2020. **Turismo: Visão e Ação**, v. 23, p. 557-578, 2021.

[2] DA SILVA, Michel Pires. **Aula 5**. Disponível em: https://ava.cefetmg.br/pluginfile.php/263776/mod_resource/content/4/Aula5.pdf. Acesso em: 22 nov. 2023.

[3] City of Los Angeles. **Los Angeles Museum Visitors**. Disponível em: https://www.kaggle.com/datasets/cityofLA/los-angeles-museum-visitors. Acesso em: 22 nov. 2023.

[4] KAMIENSKI, Arthur V.; DAMACENO, Rafael J. P. ; MENA-CHALCO, Jesús P.. **Prestígio em grafos de genealogia acadêmica: Uma proposta baseada em PageRank**. In: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING (BRASNAM), 8. , 2019, Belém. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2019 . p. 167-172. ISSN 2595-6094. DOI: https://doi.org/10.5753/brasnam.2019.6559.

[5] FERNANDES, David Augusto da Silva Paiva. **O algoritmo PageRank aplicado a redes de recomendações para seleção de recursos humanos**. Disponível em:https://repositorioaberto.uab.pt/bitstream/10400.2/7552/1/TMTSIW_DavidFernandes.pdf. Acesso em 22 nov. 2023

[6] BRIN, Sergey Mihailovich; PAGE, Lawrence Edward. **The anatomy of a large-scale hypertextual web search engine**. Disponível em: http://infolab.stanford.edu/~backrub/google.html.Acesso em: 23 nov. 2023.

[7] OLIVIERA, Marcos Vinicius; BOVOLONI, Joao Otavio; LEITE FILHO, Efraim Santana; MENEZES, Gabriel. **PAGE RANK: O FUNCIONAMENTO DA FERRAMENTA DE BUSCA DO GOOGLE**. Caderno de Graduação - Ciências Exatas e Tecnológicas - UNIT - SERGIPE, [S. l.], v. 3, n. 3, p. 73, 2016. Disponível em: https://periodicos.grupotiradentes.com/cadernoexatas/article/view/3571. Acesso em: 23 nov. 2023.

[8] HAGBERG, A. A.; SCHULT, D. A.; SWART, P. J.. **Exploring network structure, dynamics, and function using NetworkX**. In: Varoquaux, G.; Vaught, T.; Millman, J. (Eds). Anais do 7º Congresso Python em Ciência (SciPy2008). Pasadena, CA, EUA, p. 11-15, ago. 2008. Disponível em: https://conference.scipy.org/proceedings/SciPy2008/paper_2/. Acesso em: 23 nov. 2023.