Measuring Movie Script Similarity using Characters, Keywords, Locations, and Interactions

Majda Lafhel, Mohammed El Hassouni, Benjamin Renoust, Hocine Cherifi

Noname manuscript No.
(will be inserted by the editor)

Measuring Movie Script Similarity using Characters, Keywords, Locations, and Interactions

Majda Lafhel · Mohammed El Hassouni · Benjamin Renoust · Hocine Cherifi

Received: date / Accepted: date

Abstract Measuring similarity between multilayer networks is difficult, as it involves various layers and relationships that are challenging to capture using distance measures. Existing techniques have focused on comparing layers with the same number of nodes and ignoring inter-relationships. In this research, we propose a new approach for measuring the similarity between multilayer networks while considering inter-relationships and networks of various sizes. We apply this approach to multilayer movie networks composed of layers of different entities (character, keyword, and location) and inter-relationships between them. The proposed method captures intra-layer and inter-layer relationships, providing a comprehensive overview of the multilayer network. It can be used in various applications, including analyzing movie story structures and social network analysis.

Keywords Movie Script Multilayer Network · Inter layer relationships · Multilayer Graph Distance measure, Comparing movies

1 Introduction

In recent years, multilayer network analysis has achieved widespread use in various fields, including social networks, transportation networks, biological networks, and communication networks. Measuring the similarity between multilayer networks is a complex task. That is because the multilayer network consists of different entities of layers and relationships, making it challenging to define a distance measure that captures the overall multilayer network structures. Brodka et al. (2018) [1]

Majda Lafhel
E-mail: majdalafhel1@gmail.com

Mohammed El Hassouni
E-mail: mohamed.elhassouni@gmail.com

Benjamin Renoust
E-mail: renoust@gmail.com

Hocine Cherifi
E-mail: hocine.cherifi@gmail.com

have proposed a property matrix that represents a multiplex network. The property matrix maps layers and nodes into structures. Brodka et al. have used three methods to compare multiplex networks: aggregations(min, max, entropy), layer distributions(Jensen-Shannon Divergence), and similarity functions(Jaccard, cosine, correlation). Giordano et al. (2019) [2] used factorial methods for quantifying multiplex networks visually. Ghawi et al. (2022) [3] have used community detection to quantify the similarity between multilayer networks.
In previous works [4][5], we investigated Network Portrait Divergence [6] and Laplacian Spectra Descriptor [7] to compare the similarity between movie stories. We extracted for each movie a multilayer network [8] composed of three layer entities (character, keyword, and location). We ignored inter-relationships and compared monolayers of the same entities.
This research aims to quantify the similarity between movies. There have been multiple approaches to quantifying visual content [9–12, 11, 13–16]. Here, we consider multilayer network movies with inter-relationships between layers. To the best of our knowledge, there is currently no approach for measuring the similarity between multilayer networks considering inter-relationships. Analyzing the structure of interlayer relationships provides extra information and a comprehensive overview of the multilayer network. Moreover, previous studies have focused on comparing layers with the same number of nodes. Multilayer movie networks consist of layers of different sizes, making finding an appropriate measure challenging. We propose an approach that captures intralayer and intralayer relationships in the multilayer network, considering networks of various sizes.

2 Methodology

A graph $\mathcal{G}$ is a set of nodes $\mathcal{N}$ connected by edges $\mathcal{E}$. Based on this property, we consider nodes of the same entity linked by intra-relationships as graphs $\mathcal{G}_{intra}$ and nodes of different entities connected by inter-relationships as graphs $\mathcal{G}_{inter}$. We work on multilayer network movie scripts with three entities (character, keyword, and location). So, the multilayer network includes six types of networks: $\mathcal{G}_{intra_{CC}}$ is the character graph, $\mathcal{G}_{intra_{KK}}$ is the keyword graph,$\mathcal{G}_{intra_{LL}}$ is the location graph, $\mathcal{G}_{inter_{CK}}$ consists of inter-relationships connecting character and keywords, $\mathcal{G}_{inter_{KL}}$ consists of inter-relationships connecting keyword and location nodes, and $\mathcal{G}_{inter_{KL}}$ consists of inter-relationships connecting keyword and location nodes.

The proposed algorithm (Algorithm 1) maps $\mathcal{G}_{intra}$, $\mathcal{G}_{inter}$, and network features into one matrix $\mathcal{P}$ as follows.

- (i) Six rows, where the first three rows represent the three intralayers ($\mathcal{G}_{intra_{CC}}$, $\mathcal{G}_{intra_{KK}}$, and $\mathcal{G}_{intra_{LL}}$), and the last three rows represent the three interlayers ($\mathcal{G}_{inter_{CK}}$, $\mathcal{G}_{inter_{KL}}$, and $\mathcal{G}_{inter_{CL}}$).
- (ii) Six columns, each one represents a network features: max degree, max centrality, density, adjacency, Laplacian, and network portrait.
- (iii) Each cell c_{ij} encodes a network feature j of the network type i.

Algorithm 1 Matrix property extraction

input: $\mathcal{G}_{intra_{CC}}$, $\mathcal{G}_{intra_{KK}}$, $\mathcal{G}_{intra_{LL}}$, $\mathcal{G}_{inter_{CK}}$, $\mathcal{G}_{inter_{KL}}$, $\mathcal{G}_{inter_{CL}}$
output: matrix property $\mathcal{P}$

for i **in** $\mathcal{G}_{intra_{CC}}$, $\mathcal{G}_{intra_{KK}}$, $\mathcal{G}_{intra_{LL}}$, $\mathcal{G}_{inter_{CK}}$, $\mathcal{G}_{inter_{KL}}$, $\mathcal{G}_{inter_{CL}}$ **do**
  $\mathcal{D} \leftarrow$ max((deg(1), deg(2), ..., deg($\mathcal{N}$))) //return the max node degree of i.
  $\mathcal{BC} \leftarrow \max(\sum_{s \neq t \in V} \frac{\sigma_{st}(\mathcal{N})}{\sigma_{st}})$ //return the max node betweeness centrality of i
  //σ_{st}: total shortest paths passing from a node s to a node t
  //$\sigma_{st}(\mathcal{N})$: total number of σ_{st} that passing through a node n
  $Dens \leftarrow \mathcal{E}/(\mathcal{N} * (\mathcal{N} - 1))$ //return density of i
  $\mathcal{A} \leftarrow Extract\_Adjacency\_matrix(i)$
  $\mathcal{L} \leftarrow Extract\_Laplacian\_matrix(i)$
  $\mathcal{B} \leftarrow Extract\_NetworkPortrait\_matrix(i)$
  $s_A \leftarrow sum(eigenvalues(\mathcal{A}))$
  $s_L \leftarrow sum(eigenvalues(\mathcal{L}))$
  $s_B \leftarrow sum(\mathcal{B})$
  $v_i \leftarrow [\mathcal{D}, \mathcal{BC}, Dens, s_{\mathcal{A}}, s_{\mathcal{L}}, s_{\mathcal{B}}]$
end for
$\mathcal{P} \leftarrow [v_{\mathcal{G}_{intra_{CC}}}, v_{\mathcal{G}_{intra_{KK}}}, v_{\mathcal{G}_{intra_{LL}}}, v_{\mathcal{G}_{inter_{CK}}}, v_{\mathcal{G}_{inter_{KL}}}, v_{\mathcal{G}_{inter_{CL}}}]$

Consider a pair of multilayer network movies M and M'. In the first step, we extract property matrices $\mathcal{P}$ from M and $\mathcal{P}'$ from M'. Second, we flatten matrices $\mathcal{P}$ and $\mathcal{P}'$ to vectors $\hat{v}$ and $\hat{v'}$. Then, we compute the distance between $\hat{v}$ and $\hat{v'}$ using the Euclidean Distance.

$$\hat{D} = \sqrt{\sum_{i=\mathcal{G}_{intra_{CC}}}^{\mathcal{G}_{inter_{CL}}} (\lambda_{\hat{v}_i} - \lambda'_{\hat{v'}_i})^2} \tag{1}$$

3 Experimental Results

We performed experiments using movie scripts from various categories. In our previous work, it appears that the romance films were the more challenging. Therefore we concentrate on these movies. We compare: Titanic (1997), episode I of Twilight (2008), and episode II of Twilight (2009). For each movie, we extracted three layers (character, keyword, location), intra-relationships and inter-relationships. We collected ground-truth data by inviting a group of individuals to rank the similarity between romance movies. Based on the evaluation, we obtained the following ranking: Episodes I and II of Twilight are in the first rank, Titanic and I of Twilight in the second, and episodes II and three also in the second.
To illustrate the efficiency of the proposed method in quantifying the similarity between romance movies (Titanic, episodes I and II of Twilight) we compare the obtained results (Table 2) to those of previous studies (Table 1).
In a previous investigation (Table 1), the Network Portrait Divergence revealed the similarity between character layers, and the Network Laplacian Spectra detected the similarity between location layers. But, no measure indicated the similarity between keyword layers.

Table 1 Checklist table for similarity between romance movies

Measures/Layers	Character	Keyword	Location
NetLSD	✗	✗	✗
NetMF	✗	✗	✗
D-measure	✗	✗	✗
Network Portrait Divergence	✓	✗	✗
Laplacien Spectra	✗	✗	✓

According to Table 2, the distance between episodes I and II of Twilight (272.93) is the smallest. That means episodes I and II are the most similar. Indeed, the ground-truth data shows episodes I and II of Twilight are in the first rank. On the other hand, Titanic is closer to episode II of Twilight (403.52) than Episode I (450.24). However, according to the ground-truth data, both pairs of movies are second, which reveals how Titanic is far from episodes I and II of Twilight.

Table 2 Distance between romance movies using the proposed method

Romance movies	Distance
episode I of Twilight & episode II of Twilight	272.93
Titanic & episode I of Twilight	450.24
Titanic & episode II of Twilight	403.52

In brief, the proposed method revealed the high similarity between episodes I and II of Twilight and the distance between Titanic compared to both movies. In contrast, in the previous research, no measure revealed the similarity between keyword layers. Furthermore, the time complexity of the proposed technique is much smaller than the prior one. That is because we compare the overall multilayers at one time.

References

1. P. Bródka, A. Chmiel, M. Magnani, G. Ragozini, Royal Society open science **5**(8), 171747 (2018)
2. G. Giordano, G. Ragozini, M.P. Vitale, Social Networks **59**, 154 (2019)
3. R. Ghawi, J. Pfeffer, Social Networks **68**, 1 (2022)
4. M. Lafhel, H. Cherifi, B. Renoust, M. El Hassouni, Y. Mourchid, in *International Conference on Complex Networks and Their Applications* (Springer, 2020), pp. 284–295
5. M. Lafhel, L. Abrouk, H. Cherifi, M. El Hassouni, in *2022 IEEE Workshop on Complexity in Engineering (COMPENG)* (IEEE, 2022), pp. 1–5
6. J.P. Bagrow, E.M. Bollt, Applied Network Science **4**(1), 1 (2019)
7. A. Tsitsulin, D. Mottin, P. Karras, A. Bronstein, E. Müller, in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2018), pp. 2347–2356
8. Y. Mourchid, B. Renoust, H. Cherifi, M. El Hassouni, in *International Conference on Complex Networks and their Applications* (Springer, 2018), pp. 782–796
9. S. Rital, H. Cherifi, S. Miguet, in *International Conference on Pattern Recognition and Image Analysis* (Springer, Berlin, Heidelberg, 2005), pp. 522–531
10. C. Demirkesen, H. Cherifi, in *International conference on advanced concepts for intelligent vision systems* (Springer, Berlin, Heidelberg, 2008), pp. 752–763
11. S. Rital, A. Bretto, H. Cherifi, D. Aboutajdine, in *International Symposium on VIProm-Com Video/Image Processing and Multimedia Communications* (IEEE, 2002), pp. 351–355
12. A. Lasfar, S. Mouline, D. Aboutajdine, H. Cherifi, in *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, vol. 1 (IEEE, 2000), vol. 1, pp. 1031–1034
13. M. Hassouni, H. Cherifi, D. Aboutajdine, IEEE Transactions on Image Processing **15**(3), 572 (2006)
14. R.R. Pastrana-Vidal, J.C. Gicquel, C. Colomes, H. Cherifi, Proc. 5th Int. WIAMIS (2004)
15. R.R. Pastrana-Vidal, J.C. Gicquel, J.L. Blin, H. Cherifi, in *Human Vision and Electronic Imaging XI*, vol. 6057 (SPIE, 2006), vol. 6057, pp. 276–286
16. M. Messadi, H. Cherifi, A. Bessaid, arXiv preprint arXiv:2106.04372 (2021)