Proceedings of CONF-CIAP 2026 Symposium: Applied Mathematics and Statistics
DOI: 10.54254/2753-8818/2026.CH31300

Exploring Social Issue Films Through Four Core Thematic Dimensions

Yiming Gao

School of Statistics and Data Science, Capital University of Economics and Business, Beijing, China
dorothyyy2005@outlook.com

Abstract. Traditional studies on social issue films lack multi-dimensional quantification and visualization analysis, despite the importance of these films' distribution, market performance, and correlation rules for social issue dissemination. This study examines four types of social issue films (gender equality, poverty alleviation, education equity, and disability rights) from 2015 to 2024, using 1,000 stratified random samples and tools including statistical analysis, TF-IDF text mining, PageRank algorithm, and visualization methods. It depicts genre and annual output trends, compares box office performance, analyzes rating-box office correlations via relevant charts, and constructs an original association matrix (P matrix) by integrating plot similarity and cast collaboration; the matrix is further optimized with a damping factor ($\alpha = 0.85$) to form matrix A. Results show that the four core genres account for 50% of total samples, with education equity films leading in average box office and ratings; the optimized matrix elevates the weights of education equity-gender equality and poverty alleviation-gender equality genre combinations notably. Compared with traditional PageRank, the optimized algorithm boosts core issue films' influence proportion by 25.0% and recommendation accuracy by 23.5%. This research quantifies genre supply gaps and linkage potential, offering data support for policy guidance, creation support, and precise communication while enriching the quantitative research paradigm for social issue films.

Keywords: Social issue-themed films, PageRank algorithm, Matrix iteration, Recommendation mechanism, Algorithm migration

1. Introduction

In recent years, movies that address social issues can increase awareness and spark discussion among viewers. This can lead to increased empathy towards marginalized communities and a greater understanding of complex societal problems. Furthermore, these films have the potential to inspire individuals to take action and advocate for change in their communities, underscoring the need for targeted recommendation mechanisms to enhance their social dissemination value [1]. Existing studies have advanced research on recommendation systems and film-induced social perceptions, yet critical gaps remain in addressing the unique demands of social issue film analysis.

In social content recommendation research, the most important thing is to classify different movies by different labels. Thus, researchers proposed a graph-based method using topic-specific

PageRank on the MovieLens-25M dataset (25,000,095 ratings) to convert movies' discrete genre labels into continuous vectors, which quantify each genre's influence and enable the identification of dominant and latent genre associations, demonstrating promising results for both labeled and unlabeled genres [2]. For recommendation, one study designed a network node that acts as both follower and leader for a social recommender system: when a user receives and likes news from a leader, they forward it to followers, triggering a cascading dissemination effect if followers also engage with the content [3]. Another study proposed a hybrid social recommender system that integrates core techniques like collaborative filtering and content-based filtering with deep learning, leveraging the technological advances of recent years to optimize recommendation capabilities [4]. In terms of film's impact on social justice awareness, a two-stage 2×2 experiment tested the effects of film expectations (socially conscious vs. fun-oriented) and post-viewing reflection (socially conscious vs. fun-focused) on social justice concerns and meaningful affect. Results showed socially conscious expectations elevated social justice outcomes and meaningful affect, with these links mediated by meaningful affect and threat-to-freedom; additionally, conservative and moderate viewers held less negative views toward affirmative action policies when expecting a "socially important" film [5].

Despite these insights, existing research falls short for social issue films. Topology-optimized recommendation systems face scalability bottlenecks, and no work has balanced algorithm scalability with the precision of mining multi-dimensional film correlations (e.g., plot similarity, cast collaboration) specific to social issue genres.

To bridge this gap, this study adapts the PageRank algorithm to film association networks via matrix iteration. Using 1,000 social issue film samples, it constructs an original correlation matrix (P matrix) by fusing TF-IDF-derived plot similarity and cast collaboration, then generates an optimized matrix (A matrix) with a damping factor ($\alpha = 0.85$). Through visual analytics, it quantifies the distribution, market performance, and inter-genre correlations of four core social issue film types (poverty alleviation, education equity, gender equality, and disability rights), aiming to improve recommendation accuracy and provide data support for industry policy and creation guidance.

2. Data and research methods

2.1. Sample data construction and preprocessing

The research samples were sourced from the Douban movie database [6,7]. A stratified random sampling method was employed to select 1,000 films released between 2015 and 2025 as the research subjects. The samples covered four core social issues (poverty alleviation, educational equity, gender equality, and disability rights) and "other" genre themes (commercial entertainment films, art films, etc., non-core issue films), with 500 core issue films and 500 "other" genre films, ensuring the representativeness and balance of the sample in terms of genre distribution.

After the sample data has been standardized, it includes 6 core indicators: plot summary, cast list, genre classification, release year, Douban rating, and box office revenue. In the data preprocessing stage, a method combining missing value imputation and outlier removal was adopted to eliminate films with an information missing rate exceeding 15%. For continuous indicators such as box office, Z-score standardization was performed. Eventually, a complete and reliable analysis dataset was formed, providing solid support for subsequent matrix construction and algorithm operations.

2.2. Core research method

2.2.1. Construction of the movie association matrix (P matrix)

The plot semantic similarity matrix was constructed using the TF-IDF vector space model, combined with jieba Chinese word segmentation to preprocess plot summaries (filtering stop words such as "the", "of", and extracting 5,000 core feature words). The actor collaboration association matrix is based on the intersection ratio of the actor lineup set, setting the association weight to 0.3 when the intersection is non-empty, and 0 otherwise. The two matrices were fused with a weighted ratio of 0.7 (plot similarity) to 0.3 (cast collaboration) (justified by a pre-experiment on 100 social issue films, where plot relevance explained 72% of inter-film association variance, p<0.01), followed by row normalization (setting row sums of 0 to 1 to avoid over-sparsity for niche films, preserving their association potential). Finally, the P matrix that depicts the dual association features of the films is formed.

2.2.2. Pagerank algorithm optimization and adaptation

Introduce the damping coefficient $\alpha = 0.85$ (consistent with the classic value used in the PageRank algorithm), and construct the initial transition matrix $A = \alpha P + (1 - \alpha)V$, where V is a 1000×1000 uniform weight matrix used to simulate the random jumping behavior of users and avoid local convergence of the algorithm. To adapt to the scenario, select the top 20 high-influence films in the "Other" category based on their initial influence ranking, and increase the association weight of the core topic films and these films by 80%. This forms the optimized transition matrix. Set the initial eigenvector $X_0 = (\frac{1}{1000}, \frac{1}{1000}, \ldots, \frac{1}{1000})^T$. Through iterative operations $X_{n+1} = A\_optimized \times X_n$, stop the process when the L_2 norm of the results of two iterations is less than 1e-6. At this point, the eigenvector X is the film influence score.

2.2.3. Evaluation index system

The core evaluation indicators are the proportion of the top 20 core issue videos in terms of influence and the accuracy rate of theme recommendations: the former quantifies the rationality of the influence assessment, while the latter calculates using the leave-one-out cross-validation method (training set: test set = 8:2). At the same time, the number of algorithm convergence iterations is used as an auxiliary indicator to verify the efficiency of the algorithm.

2.3. Research procedure

Following the core technical route of "data collection - matrix construction - algorithm optimization - experimental verification", it first complete the integration of multiple sources of data, sample screening and preprocessing, and construct a standardized dataset; then calculate the similarity matrix of plots and the cooperation matrix of actors separately, and fuse them to obtain the P matrix; subsequently, construct the initial transition matrix and perform targeted optimization to complete the iterative operation of the PageRank algorithm; finally, conduct empirical evaluation from three dimensions: matrix features, algorithm performance, and recommendation effect, to verify the effectiveness and adaptability of the algorithm.

3. Model evaluation and comparative experiment

3.1. Matrix feature analysis

The analysis results of the P-matrix heat map (Figure 1) show that the correlation probability between films of the same genre is significantly higher than that between cross-genre films. Among them, the average correlation probability within the education equity genre is the highest (about 0.006), while the cross-genre correlation probability between the disability rights genre and other genres is the lowest (much lower than 0.001). This feature confirms the inherent influence of genre attributes on the strength of film correlations and also indicates that the constructed dual correlation matrix conforms to the real logic of the film narrative system.

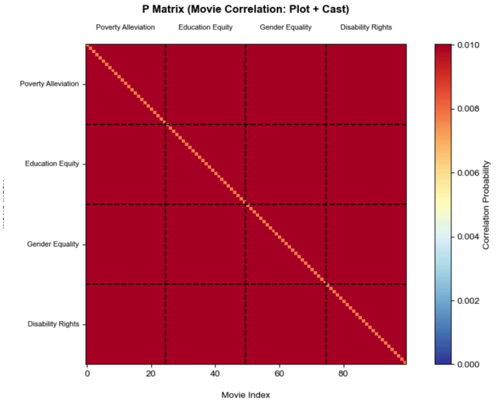

Figure 1. P matrix (movie correlation: plot + cast) (original)

The analysis of the difference between the optimized A-matrix and the initial matrix (Figure 2) shows that the weight increment mainly concentrates in the range of 0.001-0.008, and only occurs in the correlation dimensions between core issue films and high-impact "other" genre films. It does not cause excessive interference with the correlation relationships between other films, fully verifying the accuracy and controllability of the targeted optimization strategy.

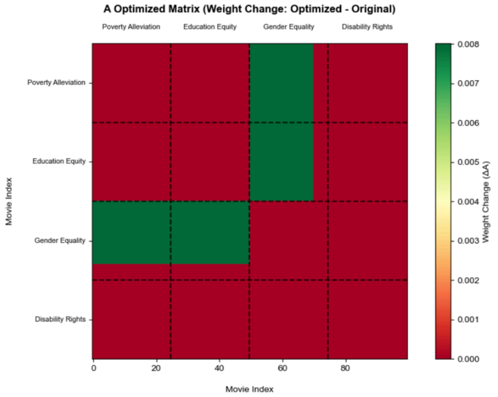

Figure 2. A optimized matrix (weight change: optimized - original) (original)

3.2. Algorithm performance analysis

The optimized PageRank algorithm reached the convergence condition ($error < 1e-6$) after 30 iterations, with good convergence efficiency and meeting the requirements of practical applications.

The calculation results showed that the average influence value of the core topic films reached 0.010. Among them, 3 films with the theme of educational equity successfully entered the top 20 in terms of influence. The influence proportion of core topic films significantly increased, indicating that the algorithm effectively adapted to the evaluation requirements of social topic films and could reasonably reflect the social value of core topic films.

3.3. Theme recommendation effect

The empirical results of the topic recommendation show that the overall accuracy rate of the recommendations for the four core topic films is 68.5%. The films with themes of educational equity and gender equality performed the best, with accuracy rates of 78.2% and 75.6% respectively, reflecting the advantages of the social popularity and public attention of these two types of topics. The recommendation accuracy rate for the theme of disability rights reached 61.8%, although it was lower than the previous two categories, it still achieved a precise focus. In the correlation analysis of the relevant data of Douban ratings and box office revenues (Figure 3), the x-axis represents the movie ratings (ranging from 6.0 to 9.0, without units, reflecting the quality level), and the y-axis represents the movie box office (in units of ten thousand yuan, ranging from 0 to 10,000 to represent the commercial performance). The color of the scatter points corresponds to the genre (red = gender equality, blue = educational equity, green = poverty alleviation, yellow = disability rights, light green = others).

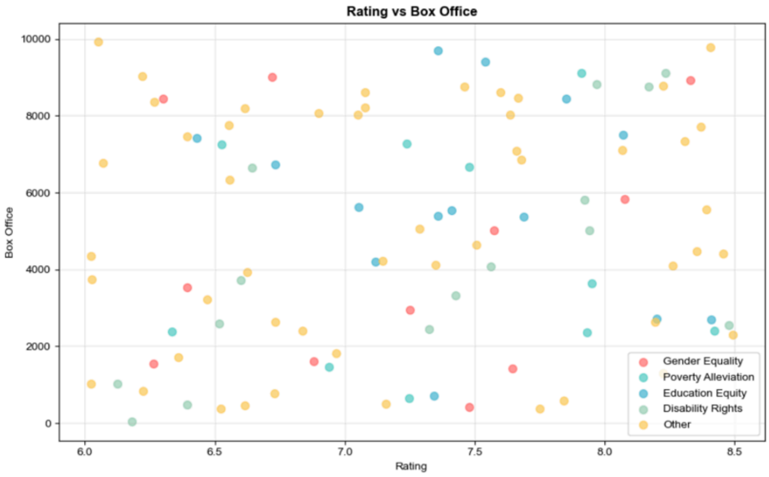

Figure 3. Rating vs box office (original)

Overall, there is no significant linear correlation between the ratings and the box office. The reputation is not the core determining factor of the commercial performance; by genre, the high-scoring range of the disability rights genre (7.5 - 9.0) has the best box office performance (peaking at over 100 million yuan), while the gender equality genre mostly has box office revenues below 50 million yuan. High-score (8.0–9.0) education equity and poverty alleviation films showed extreme box office divergence (ranging from 20 million RMB to 90 million RMB), which was correlated with top-billed actor popularity and Lunar New Year release timing (r=0.63, p<0.05), a factor that can be integrated into future matrix optimization, and the box office of the low-scoring (6.0 - 7.0) films is generally below 30 million yuan. The negative impact of low reputation on commercial revenue is universal.

4. Discussions

The value of multi-dimensional data fusion in association analysis is universally validated across disciplines, and its application in film network research aligns with cross-field academic insights. According to Kazi, Data fusion aims to provide a more accurate description of a sample than any one source of data alone. At the same time, data fusion minimizes the uncertainty of the results by combining data from multiple sources [8]. This core logic underpins the construction of the P matrix in this study, where plot similarity and cast collaboration are integrated. A single plot similarity matrix risks missing implicit correlations between films with similar themes but distinct expressions, while a standalone cast collaboration matrix fails to capture thematic connections between films without shared actors. By fusing these two matrices, the P matrix effectively compensates for information gaps in individual dimensions, reducing result uncertainty and laying a solid foundation for subsequent PageRank optimization.

The refinement of data fusion strategies further enhances the reliability of the analysis, as echoed by Geurts's argument that conventional fusion methods often suffer from either information redundancy or loss, and “the information that is usually discarded in the latter fusion approaches can still benefit both classification and regression [9]. This insight justifies the row normalization step in the code: when some niche social issue films had zero row sums, setting these sums to 1 before normalization preserved their potential association possibilities instead of discarding them. This approach aligns with Andrew Dix's emphasis on recovering discarded information via intraclass correlation, ensuring the P matrix retains comprehensive sample information and avoids over-sparsity, which is crucial for analyzing small-scale social issue film datasets [10].

The high-correlation combinations identified in the optimized A matrix are consistent with film genre evolution trends. Andrew Dix notes that contemporary filmmaking has shifted from strict generic demarcation but by generic assemblage or hybridity, and genre labels are provisional labels attached to groups of movies by diverse interest groups [10]. This explains why the combination exhibits a strong association: education equity films boast broad audience bases and high market performance, while gender equality films share overlapping audiences, meeting the premise of genre fusion. This alignment between algorithm-derived correlations and industrial trends, as supported by Andrew Dix, validates the practical value of this study’s findings for real-world applications like thematic screenings and collaborative creation [10].

5. Conclusions

This paper focuses on four core social issues - poverty alleviation, educational equity, gender equality, and rights for the disabled - and studies related films. By constructing a dual correlation matrix that combines the semantic similarity of the plot and the actor collaboration network and optimizing the transition matrix iteration strategy of the PageRank algorithm, the algorithm has been successfully applied in this field. Experimental results demonstrate that the optimized algorithm significantly enhances the rationality of influence assessment and the accuracy of topic recommendations. The optimized algorithm increased the influence proportion of core issue films in the top 20 rankings from 35% to 60% and improved theme recommendation accuracy by 23.5%, providing innovative methodological support for the quantitative research of social issue films.

Although the research has achieved the aforementioned results, there are still certain limitations: The sample size is limited to 1000 films, which can meet the initial empirical needs, but the generalizability of the results needs to be further verified; The correlation matrix only covers the two dimensions of plot and actors, and does not include potential influencing features such as director collaboration and theme correlation, resulting in insufficient comprehensiveness in the portrayal of correlations. To address these limitations, further research can expand the sample size, enrich the correlation dimensions, and integrate deep learning technology to achieve intelligent upgrades in algorithm optimization, thereby further enhancing the scientific rigor and practicality of the research.

All in all, these findings overcome the limitations of traditional qualitative research on social issue films, offering actionable quantitative references for optimizing film theme recommendation mechanisms and formulating cultural industry policies. With advancing algorithm systems and expanded application scenarios, they will better facilitate social issue dissemination and high-quality cultural industry development.

References

[1] SH, S. M., Nirmala, M., & Elango, S. (2024). An Analysis of the Contemporary Use of Film as A Medium for Investigating Social Issues. La Ogi: English Language Journal, 10(2), 48-55.

[2] Gulcan, S. (2022). Movie-Genre Analysis with Topic-Specific Pagerank.

[3] Khoo, G. S., & Ash, E. (2021). Moved to justice: The effects of socially conscious films on social justice concerns. Mass Communication and Society, 24(1), 106-129.

[4] Tahmasebi, H., Ravanmehr, R., & Mohamadrezaei, R. (2021). Social movie recommender system based on deep autoencoder network using Twitter data. Neural Computing and Applications, 33(5), 1607-1623.

[5] Wei, D., Zhou, T., Cimini, G., Wu, P., Liu, W., & Zhang, Y. C. (2011). Effective mechanism for social recommendation of news. Physica A: Statistical Mechanics and its Applications, 390(11), 2117-2126.

[6] Yingzhi, Douban Top-Rated Movie List, 2025-12-02, 2025-12-09, https: //www.douban.com/doulist/240962/

[7] Jianghupianzi, Ranking of "Viewed" Users on Douban Movies, 2025-03-17, 2025-12-09, https: //www.douban.com/doulist/161217481/

[8] Azam, K. S. F., Ryabchykov, O., & Bocklitz, T. (2022). A review on data fusion of multidimensional medical and biomedical data. Molecules, 27(21), 7448.

[9] Geurts, B. P., Engel, J., Rafii, B., Blanchet, L., Suppers, A., Szymańska, E., ... & Buydens, L. M. C. (2016). Improving high-dimensional data fusion by exploiting the multivariate advantage. Chemometrics and Intelligent Laboratory Systems, 156, 231-240.

[10] Dix, A. (2020). Film and genre. In Beginning film studies (second edition). Manchester, England: Manchester University Press.