A Literature Review of Box Office Prediction Using Online Platforms

Ⅰ.Introduction

The film industry is an industry with substantial profit potential but also significant risk. Box office revenue serves as the primary source of income for the film industry, making it an important subject for commercial research. Consequently, there has been extensive research focused on the study and prediction of box office performance. However, studies specifically using online platform data to predict movie box office revenues are relatively limited.

Our paper primarily focuses on a literature review of research in this area. Our literature review does not involve extensive mathematical comparisons; instead, it adopts a sociological and audience psychology perspective.

First, we will provide a brief overview of five key papers in this field. Following this, we will compare the similarities and differences in their research methodologies. The emphasis of this paper lies in examining the methodologies of these studies to offer a more comprehensive perspective on this domain, and to propose a new research approach along with related research questions.

Ⅱ.Articles of this field

In this field, the earliest paper we selected uses the user data from Wikipedia to predict movie box office revenues.[1] The authors utilized four parameters: the number

of page views, the number of users involved in editing, the number of human editors, and the rigor of collaboration to estimate a movie's popularity. As illustrated, these parameters have a certain degree of correlation with box office performance. Through mathematical analysis, this paper achieved an average R Square of 0.78 on a sample size of over 300, demonstrating good robustness.

Similarly, a 2017 paper focused on predicting box office revenues using user data from Twitter.[2] This study delved into three factors: the number of tweets, the rate of positive reviews, and the rate of negative reviews, to analyze box office performance. This methodology shows an overall strong relationship between box office performance and tweets related to the film. However, when the author delves into specific case studies, a film with more comments or high positive comment rate don't always indicate a more successful box office revenue.

In another paper, the authors similarly used user data from Twitter to analyze box office revenues.[3] Unlike the previous study, this paper focused more on prediction. It used data from the first week after the movie's release to make predictions and achieved an R Square of over 0.85.

On top of that, the paper published in 2016 considered motion picture, external, and audience as the sources of input variables.[4] The innovation in this study primarily lies in the mathematical and statistical methods used, which we will not elaborate in this article.

Among these papers, there is one that stands out due to its distinct approach.[5] Firstly, it utilized diverse data sources, including YouTube, Twitter, and IMDb.

Secondly, it incorporated different mathematical models to encompass a wide range of parameters. Ultimately, the algorithm identified four key attributes—popularity of the leading actor and actress, genre, and sequel status—as significant factors in predicting movie success. Surprisingly, movie reviews were not deemed important by the algorithm. According to the authors, this was likely due to the different weights assigned to various factors.

Ⅲ.Analysis of Their Methodologies

In this section, we will conduct a more in-depth and comprehensive analysis of these papers' methodologies, focusing on several categories.

Platform

These studies primarily focus on three platforms: Wikipedia, Twitter, and YouTube. The choice of platform significantly affects the research methodology. Different platforms have distinct user behaviors and audiences, resulting in significant differences in the types of data and the representative populations. For example, on Wikipedia, due to the platform's characteristics and the relative ease of accessing its API, the volume of data researchers face is much smaller compared to Twitter. However, the richness of the data is distinctly different. Wikipedia analysis can include metrics like the number of times an article is edited, the number of users editing the article, and even estimations of the article's quality. In contrast, Twitter analysis focuses more on comprehensive natural language processing. These data

focuses more on human to human interactions, as well as emotions. YouTube data is simpler and more direct, involving metrics like the number of fans of actors and the number of trailer views. On top of that, we believe that platform choice can potentially lead to statistical bias. If researchers focus only on one social media platform, data bias can occur due to the specific user demographics and structure of that platform. Therefore, we conclude that multi-platform analysis makes the research more robust.

Variables

Most articles select variables based on audience reviews. This selection depends on the platform chosen and the nature of the internet. In class, we discussed that the internet is essentially a network of nodes and links, with nodes representing users and links representing the connections between users. Due to the anonymity of the internet, it is almost impossible to study each user individually. Unlike in real life, connections between nodes on the internet are simpler, usually in the form of publicly visible text. This makes natural language processing a common tool in online platform research, not just for box office predictions. However, one of the papers we mentioned earlier does not involve natural language processing at all. Instead, it uses traces of user activity online, such as watching videos, following users, and liking tweets, to provide a different analysis approach. This approach is highly inspirational for future research.

Ⅳ.New Approach to This Field

After reading articles on box office analysis unrelated to online platform analysis, we gained a deeper understanding of this topic. Box office data largely reflects the probability of audiences willing to go to the cinema to watch a movie. This provides two research approaches.

First, we can study from a causative perspective, analyzing the factors that influence audiences to go to the cinema. For example, past articles have researched the impact of an actor's fame on box office performance[6] and the influence of promotional methods[7]. These articles analyze how these factors causes the box office revenue to increase or decrease.

Additionally, we can study from a result-oriented perspective, analyzing what outcomes arise when audiences are willing to watch a movie, such as the level of discussion about the movie. Collecting such data in the real world is very challenging and requires extensive public opinion surveys. However, with online platforms, user discussions are in the public domain, making this angle of research easier and more efficient. This has been proven true: most studies in this area analyze user discussions to research box office performance, achieving high accuracy.

However, we propose that analyzing the causes of box office success or failure through online platforms is equally necessary. This is an important yet under-explored area for the following reasons:

First, the conclusions of these studies generally do not have strong predictive power. Much of the data needs to be collected a week or even a month after the

movie's release. This limits the commercial value of these studies.

More importantly, the conclusions of these studies do not provide actionable insights for professionals in the film industry. These articles show the correlation between user behavior and box office performance but do not, and cannot, provide causal research. Film industry professionals find it difficult to understand how to design a product that the market will accept based on these studies. As the internet evolves, online platforms increasingly influence individual decisions. Utilizing the features of online platforms to better adapt to the movie market can significantly benefit the film industry.

Therefore, we propose a new research approach in this field, which consists of two steps.

First, researchers need to analyze the impact of specific factors on box office performance. These questions have typically been thoroughly researched, both mathematically and sociologically. Researchers can reach conclusions by integrating previous studies.

Second, researchers need to study how these various factors are represented and symbolized on online platforms. API calls online are more convenient and have a broader reach compared to public opinion surveys. Researchers need to analyze data from online platforms to investigate information related to the factors influencing a specific movie. Then, by applying the conclusions from the first step, they can predict the box office performance of a particular movie.

To better explain our proposal, we will illustrate with a concrete example:

Analyzing the Impact of Famous Actors on Box Office Performance

In a 2012 study, the authors clearly defined what constitutes a movie star and their impact on box office performance.[6] The authors used precise figures to describe an actor's bankability—the ability to attract box office revenue—and the quantified benefits of replacing an ordinary actor with a star. With the rise of social media, subsequent research can reanalyze a specific actor's bankability. Researchers can provide an rating of each actor’s bankability based on data such as the number of fans, tweet engagement, and other social media metrics. By combining the new rating system and the original study in 2012, researchers will be able to answer questions such as : if we replace Leonardo Dicaprio with Ryan Gosling, what will the change of our benefit be. This can provide better strategies for film industry professionals in casting and budget planning.

Similarly, many other factors, such as movie promotion strategies, can be studied. These research questions have been explored in the real world. Researchers need to connect the real world with online platforms more extensively. By leveraging the vast amount of data and the advantages of publicly available data on online platforms, more scientific and accurate conclusions can be drawn.

Ⅴ Conclusion

In summary, we believe that online platforms are an extension and reflection of real-world society. Most current research focuses on finding correlations between data from online platforms and real-world data. However, we argue that it is more important to build a bridge connecting the real world and online platforms from the perspective of understanding social computing and the mechanisms of online platforms. This makes causal analysis of data changes on online platforms possible and, conversely, helps us quantify the abstract conclusions obtained from our sociological analyses in the real world.

[1] Mestyán, M., Yasseri, T., & Kertész, J. (2013). Early prediction of movie box office success based on Wikipedia activity big data. PloS one, 8(8), e71226.

[2] Choudhery, D., & Leung, C. K. (2017, July). Social media mining: prediction of box office revenue. In Proceedings of the 21st international database engineering & applications symposium (pp. 20-29).

[3] Liu, T., Ding, X., Chen, Y., Chen, H., & Guo, M. (2016). Predicting movie box-office revenues by exploiting large-scale social media content

[4]N. Hur, P. Kang, and S. Cho. 2016. Box-office forecasting based on sentiments of movie reviews and independent subspace method. Information Sciences 372, 608-624.

[5] Apala, K. R., Jose, M., Motnam, S., Chan, C. C., Liszka, K. J., & de Gregorio, F. (2013, August). Prediction of movies box office performance using social media.

[6] Nelson, R. A., & Glotfelty, R. (2012). Movie stars and box office revenues: an empirical analysis. Journal of Cultural Economics, 36, 141-166.

[7] Brown, A. L., Camerer, C. F., & Lovallo, D. (2012). To review or not to review? Limited strategic thinking at the movie box office. *American Economic Journal: Microeconomics*, *4*(2), 1-26.