Information Analysis of Movie Genres

Janette Dauenhauer, Joneta Hockett, Joanne Mammarelli, and Michael Yarem INFO 633 – Information Visualization College of Information Science and Technology Drexel University

Abstract—The popularity of movie genres over time seems to follow cycles of ebb and flow. This paper looks at possible influences in popular culture and movie industry economics. Data is examined from the Internet Movie Database (IMDb) and The Numbers, an online movie industry information service. The data is then visualized using Microsoft Excel, Many Eyes, and Google Fusion Tables. We find that there is evidence that cycles of current events and movie profitability are linked to the production of movies within certain genres.

Index Terms— box office, genres, Google Fusion Tables, IMDb, information visualization, Many Eyes, movies, tickets

1 INTRODUCTION

The entertainment and movie industry is a multibillion dollar industry that is often a reflection of the major events and changing cultures across society. In order to better understand the entertainment industry, movies and other forms of entertainment can be categorized broadly into large groupings or genres. By classifying movies into genres such as action, comedy or drama, we can better judge the merits of a film and understand the intent of its filmmakers. Genres are key to our understanding of the story we are about to see on the large screen.

Looking at the popularity of movie genres over time, we can see cycles of ebb and flow; that is, some genres become more popular while others become less so. The reasoning behind these shifts in popularity can be due to a number of individual factors or combinations of factors. In this paper, we will investigate several factors that might account for these shifts in popularity including world events, costs to produce a movie, and revenue from a film.

To do this, we will examine genre data from the Internet Movie Database (IMDb), an online repository of movie and television information, and compare it to box office and budget figures from The Numbers, an information service for the movie industry. We will also look at possible clues in Facebook "Like" data and current events via Infoplease. We will then use the visualization tools Many Eyes and Google Fusion Tables to try to expose patterns and relationships in the data.

1.1 Literature Review

Previous studies using data from the IMDb have visualized various aspects of the movie industry, including popularity, box-office vs. DVD sales, actormovie networks and ratings. In one study, Blecksmith, Mazzola, Pellegrino and Schmidt state that "box office figures are useful to this study as a measure of a film's popularity during the time of its theatrical release." [2] The data used in this study was from the "top 50 titles of Box Office Mojo's top 100 grossing domestic films of all time". [2] They used Many Eyes from the Visual Communication Lab at IBM and Tableau software from Stanford University. Their initial goal was to find out what interesting relationships could be revealed among this movie data using various visualizations. Their conclusions were that the visualizations rendered from the data "confirms the lack of reliability that box office earnings and Oscar attention contribute to cinema popularity". [2] Their suggestions for further study include more in depth analysis of Oscar nominations vs. wins and motivations behind user ratings.

Herr, Ke, Hardy and Börner used the IMDb for their study reported at the 11th Annual Information Visualization International Conference in July 2007. They used the IMDb for three reasons:

- 1. "Most people know about and can relate to movies and actors,
- 2. The dataset has rich information on each movie and actor allowing for a wide variety of data analyses, and

3. The dataset is sufficiently clean and structured so that analysis can be done without using semantic matching techniques". [5]

The main goal of their report was "to give a global overview of the entire movie and actor space". [5] They discussed the sheer numbers of the data that was required to accomplish this goal and thus render it so it could be used and "overlay the co-actor network directly on top of the movies". [5] The co-actor network is made up of "actors that appear in a movie together". [5] This visualization was planned to be plotted in a 36" by 40" area, but with the amount of movies used, the size of the ending product was 73" wide. While analyzing the data, they found that the most interesting part of the actor layer was that "most of the actors are tightly packed into one cluster". [5] They did not further explore this anomaly but conjectured that it might increase someone's chances of being nominated or receiving an Academy Award for best actor/actress by their working with actors in the cluster. Although this project was plotted on paper, future work could be projected and interactively manipulated to help highlight interesting parts of the visualization.

Another visualization and analysis using the IMDb was conducted within a case study using the "large and complex temporal multivariate networks" derived from it. [1] The large networks that are discussed in this paper include webgraphs, social networks and biological networks. As discussed in this course (INFO 633) and emphasized in this paper about the IMDb, good visualization helps to understand data, especially complex data. Ahmed, Batagelj and Fu state that "good visualisation reveals the hidden structure of the networks and amplifies human understanding, thus leading to new insights, new findings and possible predictions for the future". [1] They add that "analysis tools for networks are not useful without visualization," and "visualization tools are not useful unless they are linked to analysis". [1] The paper discussed the visualization of the IMDb dataset in three different ways: visualization for large bipartite graphs, visual analysis based on the Kevin-Bacon number, and a visual analysis of a galaxy metaphor visualization of a temporal two mode actormovie network. The authors found some problems with the dataset, one being that the years of some of the movies had been recorded incorrectly. This was a problem with those movies that were remakes of

older movies. Since the case study dealt with such a huge amount of data, subgraphs were defined to help reduce visual complexity.

Mark Lee, in his annual statistical review of the Top 250 movies list, uses the data from the IMDb to discuss the problems with the way that these movies are rated and then ranked. He states that the "real value of this data is in its contribution to the 17-yearlong trend analysis." [9] Another Internet blog uses PageRanking to filter the data in order to discuss link analysis of the database movie connections.

The literature demonstrates how the IMDb has the potential to be reviewed in many different ways and analyzed far into the future.

2 TOOLS

As previously stated, the main source of movie genre data will be the IMDb. The IMDb is a large online database of information that contains a rich amount of information related to films, including information related to actors, directors, genre, production costs, and revenues. The IMDb was launched nearly 25 years ago in 1990. [8] The IMDb now contains nearly 3 million titles and information related to those titles. The data from the IMDb will be pulled into a tabular format using Microsoft Excel, which will allow for structured analysis of the information using several different tools. Budget and box office data will be pulled from The Numbers, an online information resource about and for the movie industry. [10] Our primary visualization tools will be IBM's Many Eyes, a free, web-based visualization software application [6], and Google Fusion Tables, a data management web application that affords several data visualization options. [4] Several types of information visualizations will be utilized via these tools to allow us to examine the rises and falls of genre popularity over time, and other influences related to the movies. Genre definitions are included in Appendix 1.

3 METHODS

The IMDb will provide the primary source of information regarding the movies we will be profiling focusing on year of production, title, genre, revenue, and cost to produce. There will be three pulls of information spaced a decade apart from 1990, 2000, and 2010.

In addition to this information, we will examine major world events happening around our data pulls to examine what, if any, influence these events have on the movies being produced and their popularity.

Data regarding production budget and gross will come from The Numbers, a web site by Nash Information Services, LLC devoted to publishing movie financial data. [10] Nash Information Services is a consulting firm specializing in providing services and information on the movie industry.

Visualizations will come from Many Eyes and Google Fusion Tables.

3.1 Data

3.1.1 IMDb

As mentioned above, the IMDb database is a large online database of information. The makers of the site allow for the downloading of the text files, which contain information from the database. The genres.list file was chosen because it identified genres associated with the movies.

Challenges were encountered when dealing with the genres.list text file. First, the compressed size of the file is 13000 KB. To meet the requirements of Many Eyes, the file had to be broken into smaller portions. Second, the file is a text file with no primary key or distinct fields upon which to search. The lines are composed of the name of the movie followed by the year it was released and the genre. There can be numerous lines if a movie belongs to more than one genre. Below is an example of the contents of the file:

Dance for Life (2012)	Documentary
Dance for Life (2012)	Drama
Dance for Life (2012)	Short

For this exercise, the Microsoft Excel and the 'Data -Text to Columns" option was repeatedly used to break the file into movies released in 1990, movies released in 2000, and movies released in 2010. Furthermore, the genres.list file contained not only movie data, but television, music and other formats. These extraneous data had to be removed from our resulting files so that that focus would be on movies and topics associated with films.

3.1.2 The Numbers

Data regarding movie budgets and gross revenues were copied from The Numbers and pasted into an Excel workbook. The data was then sorted by date so that movie information in the years 1990, 2000, and 2010 could be easily extracted. Genre information was then added from previously generated IMDb data tables. Movies categorized with more than one genre were listed multiple times, once for each genre. The Numbers also categorizes movies with genres, but these did not correspond exactly to the genres provided by the IMDb. We decided to use movie information only for those movies where the genres concurred between the two sources. This meant that the following genres were eliminated from the total set: Biography, Talk-Show, Adult, News, Sport, Romance, History, Reality-TV, Short, Game-Show, Mystery, and Music. The following remaining genres were analyzed for this effort: Action, Adventure, Animation, Comedy, Crime, Documentary, Drama, Family, Fantasy, Horror, Musical, Sci-Fi, Thriller, War, and Western.

The resulting data set for financial analysis was considerably smaller than the original IMDb data pull, as only movies in 1990, 2000, and 2010 whose budget and worldwide gross figures are known are listed. The Numbers authors acknowledge that there are gaps in the data, but that the figures are as accurate and complete as possible. [10]

3.1.3 Infoplease

In addition to this information, we examined major world events happening around our data pulls to examine what, if any, influence these events may have had on the movies being produced and their popularity. The three-year slices of information that we chose for this effort (1990, 2000, 2010) were used to find major world events happening in each one of those years. The website, Infoplease (www.infoplease.com), allows the user to search for historical events based on a wide range of variables. The three-year slices were input to the search criteria on the page which produced a free text form version of world events throughout that selected time slice.

The raw information from the site gives information in mostly single sentence, headline type of description of the events followed by a month and date of the event. This is a sample of the output from the site:

First Lady Hillary Clinton officially enters N.Y. Senate race (Feb. 6).

In order to avoid the dates within the year from appearing in our generated visualization, the dates were scrubbed from the output. The month- and dayrelated information, when included in a visualization, caused the result to be dominated by that information that would not have an impact on the result set.

Table 1. Tallies of Movies in IMDb in Each Genre for the Years 1990, 2000, and 2010

	1990		2000		2010			
Genre	No. of Movies	% of Total	No. of Movies	% of Total	% Change from 1990	No. of Movies	% of Total	% Change from 2000
Action	548	9.77%	534	5.73%	-41.40%	2048	4.46%	-22.20%
Adventure	152	2.71%	206	2.21%	-18.50%	1309	2.85%	29%
Animation	289	5.15%	602	6.46%	25.40%	1756	3.82%	-40.90%
Comedy	992	17.69%	1777	19.08%	7.90%	8793	19.15%	0.40%
Crime	314	5.60%	422	4.53%	-19.10%	1385	3.02%	-33.30%
Documentary	663	11.82%	1570	16.86%	42.60%	6734	14.67%	-13%
Drama	1438	25.65%	2365	25.39%	-1.00%	13052	28.43%	12%
Family	239	4.26%	396	4.25%	-0.20%	2030	4.42%	4%
Fantasy	149	2.66%	252	2.71%	1.90%	1703	3.71%	36.90%
Horror	211	3.76%	276	2.96%	-21.30%	2135	4.65%	57.10%
Musical	61	1.09%	100	1.07%	-1.80%	395	0.86%	-19.60%
Sci-Fi	117	2.09%	216	2.32%	11.00%	1218	2.65%	14.20%
Thriller	344	6.14%	500	5.37%	-12.50%	2621	5.71%	5.40%
War	71	1.27%	78	0.84%	-33.90%	531	1.16%	38.10%
Western	19	0.34%	19	0.20%	-0.41%	207	0.45%	125%

3.1.4 Facebook

Additional analysis included mapping the movie names from "the-numbers" with the number of "Likes" in Facebook (<u>www.facebook.com</u>). A Facebook "Like" allows the user to identify whether the consumer liked the movie or not ("Like"). An actual database with the "Like" information was not found, so we manually searched for the movie page within Facebook and captured the information. The data was saved in a tabular, Microsoft Excel format.

3.2 Visualization Tools

3.2.1 Many Eyes

IBM Many Eyes' website (<u>http://www-</u>958.ibm.com/software/analytics/manyeyes/) offers a variety of visualizations. The site requires registration in order to upload data sets and once signed in will allow you to enter your data information through a

link labeled "participate". Any data uploaded to the site will be visible to every audience with no ability to keep information confidential. As our information was already publicly available, we did not have to take any additional precautions to mask or obscure the data we were using, although this would be a large consideration for any private or identifiable information that was going to be used through this site.

After the data is uploaded the site allows the information to be visualized through a wide range of tools. Some of the tools are better suited for different types of data, such as formatted (table) information while others are better suited for free form data.

Using the stripped out data, Many Eyes was employed to create the visualization in the form of Bubble Charts which easily illustrated the size of the various genres for each corresponding decade (Appendix 2). The expectation was to render comparable visualizations for the three decades selected using 27 different genre types. However,

since the size of data being uploaded into a ManyEyes visualization is limited to 5 MB, the year 2010 had to be divided to make the visualizations possible. It was determined to separate the file alphabetically: the first part being Action to Game-Show genres, and the second being History to Western genres. The resulting data in the two files was not distributed evenly, but each was able to be uploaded into Many Eyes to render the bubble chart needed for analysis.

After the 2010 Movie genre data was uploaded and the visualizations were created (2010 Movie Genres Pt 1 and Pt 2), it was determined that in order to maintain consistency in the analysis, both the 1990 and 2000 decades needed to be divided. Each of these decades were split alphabetically in the same manner; Part 1 – Action to Game-Show genres, and Part 2 – History to Western genres. Although the files for these two decades could have been uploaded as a whole, the division was made to make comparing the data more effective.

The data was then pared down to ensure that it only included genre information about movies and anything other than movies was deleted, such as music, short films (which were usually television productions) and other television shows so labeled as television. The data then included only 15 movie genres which reduced the size of the data to be manipulated to less than the 5MB allowed in Many Eves. Bubble Charts were created to compare the count of each movie with the genre, but then it was determined that there was a big difference (in actual count) between 2010 and the other two decades. Using the raw data was not going to be effective in making a comparison. It was then determined that comparisons should be made with percentages within each genre per decade.

Another change for comparing data was made in that the best use of the information came from the pivot tables which included the movie genre, count of movie titles within each genre and each decade, as well as the selected decade. Each genre was reduced to a percentage using the total number of movies produced in that decade divided by the count within the movie genre. Percentages were applied to all genres within each decade.

3.2.2 Google Fusion Tables

Separate Excel worksheets of data from the Numbers for each of the three years examined—1990, 2000, and 2010--were saved in CSV format and uploaded to Google Fusion Tables. Google Fusion Tables was used because it lent itself to certain chart and network visualizations, namely stacked bar charts and network diagrams. The "summarize" checkbox was selected to show average budgets and total gross figures for overall genres. The data summaries are included in Appendix 3.

4 RESULTS

Our first data visualization was a layered line graph using Excel (Figure 1). Each pivot table was charted in an Excel spreadsheet as a line graph in a given color (blue for 1990, green for 2000, and red for 2010). The line graphs were layered over each other to enable the comparison between the three decades. The visualization showed the variations between genres in the decade. The differences between the decades of 1990 and 2000 are considerably less than those for 2010 in every genre, except "Musical" and "Western" which shows that counts for those two are relatively the same.



Figure 1. Layered line graph showing the number of movies made in each genre for the years 1990, 2000, and 2010.



Figure 2. Treemap for comparison visualization comparing percentages of genres for the years 1990 and 2010. Genres that saw fewer movies produced in 2010 as compared to 1990 were Action, Crime, Animation, Musical, Thriller, and War. The largest genre increases were Horror, Fantasy, Sci-Fi, Documentary, and Western.

A series of treemaps for comparison were generated in Many Eyes to show the overall distribution of genres for each year, one of which is shown in Figure 2 (see Appendix 4 for visualizations comparing 1990 to 2000, and 2000 to 2010). We chose the treemap for comparison because it allowed comparison between two datasets, in this case data for 1990 and 2010. The genres in blue show a decline in the number of movies made in a particular genre, while movies in orange show an increase.

As previously stated, because the total number of movies in all genres grew exponentially from 1990 to 2010, it did not make sense to use actual numbers for these comparisons. Rather, we used percentages of the totals. These figures are shown in Table 1. Most notable were declines in Action, Crime and Animation movies, while Horror, Fantasy, and Sci-Fi movies were on the rise. Dramas and Comedies were consistently a sizable portion of the total movies made over the years examined.

We also visualized budget and worldwide gross information for each genre (Figure 3a-c). The stacked bar charts show the proportion of each budget compared to the overall gross for movies in each genre. For 1990, the genres with the highest return on investment are Family, Fantasy and Western. Genres with the largest budgets are Action, Thriller, and Crime. Genres with the highest return on investment in 2000 are Adventure, War and Action. In 2010, the most successful movies were Fantasy, Adventure and Musical, which were also the movie genres with the largest budgets.

March 16, 2014







Figure 3a-c. Average budget and worldwide gross figures are summarized by genre, with budget shown as proportionate to worldwide gross. Note some genres are not shown for 1990 (a) because budget and gross figures were not available.

It should be noted that there is overlap between genres. Some movies, for example, are categorized as both comedies and dramas. A few movies were given as many as 6 genre categories in the IMDb dataset.

Figures 4a-c illustrate how movies cluster around certain genres.

Figure 4a-c. Network visualizations showing relationships between movie genres. Genres are indicated with yellow nodes, while blue nodes correspond to movie titles. Node size corresponds to total gross figures. 1990 featured the fewest number of nodes





Figure 5a-c. Word clouds of current events for the years 1990 (a), 2000 (b), and 2010 (c).

The free form text files retrieved from Infoplease were used to create three visualizations (Figures 5a-c) using Many Eyes. In order to generate the visualization, the files had to be each selected and then pasted into the provided input box. The tool will then automatically try to interpret your input (the tool was able to process the text files properly) and then requires the data to be titled (which is mandatory) and then allows for other information, which can include a larger description, to be attached to your data set.

After the data is uploaded the site allows the information to be visualized through a wide range of tools. Some of the tools are better suited for different types of data, such as formatted (table) information while others are better suited for free form data.

Because these data sets were free form, our visualization choices were limited. After experimenting with the phrase net, word tree, and tag cloud choices, the word cloud visualization was chosen for this analysis. The reason for this choice is that the information that we had downloaded were collections of non-linked sentences where phrases or connecting words would not be as valuable to us but individual words were more important to our analysis.

What we expected to see with this effort were large concepts, people, and countries that are commonly seen in news events. In looking at an overview of major world events, smaller news stories would be pushed out of the main focus and other stories would end up possibly repeating or linked via single words. We were very interested in what the resulting cloud would look like and what similarities would be seen in the course of the three resulting visualizations.

Finally, we graphed a comparison of Facebook likes against overall movie profits (Figure 6) to observe whether any correlation exists between the "like" measure of popularity and overall genre profitability. We found that there is no distinct pattern between the number of likes and the profitability of the movies as shown by the high like columns for unprofitable movies



Unprofitable Movies & Facebook Likes Profitable Movies & Facebook Likes Figure 6. Graph on left shows the number of likes for unprofitable movies in a particular genre, while the graph on the right shows the number of likes for profitable movies in a particular genre.

The genre data shows definite trending among certain movie genres. Comedies and dramas are consistently the bulk of the movies made in each of the three years examined. This is easily seen in the bubble charts in Appendix 2. In terms of net profit, comedies and dramas are consistently the most profitable, with the lowest budgets and highest gross. But when looking at the percentages of the whole shown in the treemap for comparisons in Figure 2, it appears that action movies are experiencing a decline. Looking at the budget and profitability data, this fact seems inconsistent. The actual profit from action movies has risen from 1990 to 2010. And while the number of documentaries made has increased, the net profit has remained relatively low from 2000 to 2010 (1990 did not have documentary data). If Facebook "likes" are a measure of popularity, the fact that some unprofitable movies can get a lot of likes while some profitable movies get none only deepens the mystery surrounding the rise and fall of certain genres.

If popularity alone is not enough to drive movie profitability, then some other forces may be at work. The word clouds for each of the years considered here show a subtle shift in events that may shape public attitudes when going to the movies. In 1990, the main thrust of events surrounded events in Europe (e.g., the imminent dissolution of the Soviet Union), and in 2000 most of the focus was on the presidential election of that year. Both word clouds represent relatively benign events from a domestic point of view. By 2010 (the post 9/11 world), events had shifted in focus to events tied to wars that directly involved the United States. It is possible that this shift in public concern could have negatively impacted the profitability of war movies, which showed a net loss in 2010 of 2.15%, and the decline in the genre overall.

Given all this, there is still the tangled web of movie clusters to consider. While genres such as dramas and comedies remain the most profitable, movie genres closely associated with these genres also seem to inherit from their influence. For example, Thrillers are frequently co-categorized with Action movies, yet Thrillers have shown a large increase in profits as of 2010, despite an overall decline in production from 1990. The Thriller may benefit, however, from the relatively lower budget required to make them.

Also clustered together frequently are Comedy, Family, Fantasy, and Animation movies. These movies have all enjoyed an upswing in both production and profitability from 1990 to 2000 to 2010. It could be postulated from the network cluster visualizations that in making an animation movie, making a family-friendly comedy animated movie would be a better financial bet than an animated drama.

6 CONCLUSION

The visualizations generated in this exercise revealed that movie popularity and movie profitability are not necessarily linked. Overall profit seems to be a better indicator of what kinds of movies will be produced. It also appears that current events can have an impact on public tastes when it comes to purchasing a movie ticket.

The movie industry is a thriving business and the IMDb will continue to add new data as long as movies are produced, whether they are released in the theatres, on media or via video streaming. There are many parts to the industry as well, such as the production, the direction, the awards, the actors and the generated revenues. All of these still can be analyzed through information visualization processes that filter the information needed and render it via simple line graphs or complex 3-D renderings.

REFERENCES

- [1] Ahmed, A., Batagelj, V., & Fu, X. (2007, February). Visualisation and analysis of the Internet Movie Database. Retrieved from <u>http://sydney.edu.au/engineering/it/~dmerrick/papers/Ahm</u> <u>edEtAl2007.pdf</u>
- [2] Blecksmith, A., Mazzola, M., Pellegrino, K., & Schmidt, J. (2008, March). An exploration of great cinema using information visualization platforms. Retrieved from <u>http://cluster.ischool.drexel.edu/~cchen/courses/INFO633/</u>07-08/g4.pdf
- [3] Facebook. (2014). Like. Retrieved from https://www.facebook.com/help/like
- [4] Google. (n.d.). Google Fusion Tables. http://tables.googlelabs.com/

- [5] Herr, B. W., Ke, W., Hardy, E., & Börner, K. (2007, July). Movies and actors: Mapping the Internet Movie Database. Retrieved from <u>http://nwb.cns.iu.edu/papers/2007-herr-movieact.pdf</u>
- [6] IBM Corporation. (2010). Many Eyes. <u>http://www-958.ibm.com/software/analytics/manyeyes/</u>
- [7] Infoplease. (2014). Retrieved from http://www.infoplease.com http://www.infoplease.com/ipa/A0005272.html (1990 data) http://www.infoplease.com/ipa/A0880912.html (2000 data) http://www.infoplease.com/year/2010.html (2010 data)
- [8] Internet Movie Database. (2014). What is IMDb? Retrieved from <u>http://www.imdb.com/help/show_leaf?about&ref =hlp_br_ws</u>
- [9] Lee, M. (2012, October 08). IMDb top 250 movies list analysis, 5th edition. Retrieved from <u>http://www.overthinkingit.com/2012/10/08/imdb-top-250-movies.-5th-edition/</u>
- [10] Nash Information Services, LLC. (2013). Movie budgets. Retrieved from <u>http://www.the-numbers.com/new/movie/budgets/all</u>