An Exploration of Great Cinema using Information Visualization Platforms (March 2008)

Anne Blecksmith, Missy Mazzola, Katherine Pellegrino and Jessica Schmidt

The iSchool, Drexel University
Distance Learning Program
Info633: Information Visualization
Professor Chaomei Chen

Abstract -- This paper uses the information visualization platforms of Many Eyes and Tableau to visually assess and analyze what elements make a film more popular than another. Data compiled includes but is not limited to: user ratings from the Internet Movie Database, box office business, release date, film genre, Oscar nominations and Oscar wins. The list of films explored consists of the IMDb's top 100 rated films as of March 2008.

***Index Terms*— film, great cinema, IMDb, information visualization, Many Eyes, movies, Oscars, ratings, Tableau**

I. Introduction

Cinema has flourished in America since the rise of Hollywood in the early 20th century. As cinema has embedded itself in our culture, it has become an easy topic of social conversation, as most people can easily talk for hours about their favorites. Over time, certain films have transcended personal opinion to become commonly held as truly great films. What makes these films greater than others? What is it about them that make them more popular? By compiling basic film data from the Internet Movie Database's Top Rated 100 films and translating it into a visual format using the information visualization tools of the platforms known as Many Eyes and Tableau, these questions start to develop answers in the form of visual patterns.

First, a short introduction to the origins of our data is needed. Internet Movie Database, or IMDb, is an incredibly extensive database that was launched in 1990 and later acquired by popular internet retailer Amazon.com in 1998 [1] Besides offering detailed information about any major film ever made, the IMDb also has a user rating system wherein movies can be rated on a scale of 1 (awful) to 10 (excellent). To avoid ballot stuffing, the following weighted formula is used for calculating the Top Rated 250 Titles:

$(WR) = (v\ )(v+m))\ HR + (m\ )(v+m))\ HC$
where:
WR = weighted average
R = average for the movie = (Rating)
v = number of votes per movie = (votes)
m = minimum votes required to be listed in the Top 250 (currently 1300)

C = the mean vote across the whole report (currently 6.7) for the Top 250, only votes from regular voters are considered.

For the scope of this study, only the first 100 titles of the IMDb Top 250 are utilized.

Box office data used in this study was compiled from the top 50 titles of Box Office Mojo's top 100 grossing domestic films of all time [2]. Figures have been adjusted for inflation. Founded in 1999, Box Office Mojo is a public website which describes itself as an "online movie publication and box office reporting service." Most figures are available to the public; however, the site offers a "Premier Pass" subscription which allows users to make use of the full version of the site for a fee. Although not always exact, box office figures are useful to this study as a measure of a film's popularity during the time of its theatrical release.

The next section presents an overview of the visualization tools used in this study.

II. Visualization tools

2.1. Many Eyes

Launched online in 2007, Many Eyes is at once a public website "for shared visualization and discovery" and a research experiment of the Visual Communication Lab which is part of the Collaborative User Experience Research Group at the IBM Watson Research Center [4]. As a result, visitors to the site not only share their data and visualizations with each other—they are also sharing their experiences with the Visual Communication Lab.

In order to create visualizations, users upload their data to Many Eyes as tab-delimited text files: i.e. values must be separated by tab characters and not spaces or commas. Users may also choose to use pre-existing data sets saved on the site either alone or combined with new data.

Once data has been uploaded to the site, users may choose from an established taxonomy of fourteen display options or "visualizations" that include maps, line graphs, stack graphs, bar charts, block histograms, bubble charts, matrix charts, scatterplots, network diagrams, and treemaps. Visualizations are saved to the site and are thus searchable and browseable by all visitors to the site. In some instances, Internet search engines will also retrieve visualizations created with Many Eyes.

Figure 1 -- Visualizations created using Many Eyes

According to its founders, Many Eyes was modeled after popular social networking and file sharing sites, such as myspace.com and flickr.com, designed for social interaction between users [3]. Data sets and visualizations have linked discussion areas where other users are

free to write comments. Visualizations may also be given rankings from individual users on the site. Consequently, visualizations may be sorted by user rankings.

Many Eyes' users may also create "topic hubs" where visualizations may be linked to others based on a common theme, category or subject. Many Eyes also sponsors a blog on its site and encourages users to link into the site from their own independent websites.

At present, over 10,500 individual visualizations have been created and archived on the site. Since its launch, users have created visualizations at the rate of about 700 per month. The site appears to be achieving the Visual Communication Lab's twofold goal: to provide a forum where the public may have access to visualizations tools and discuss the results while participating in a "case study" developed and observed by information visualization researchers [7].

2.2 Tableau Software

What once began as a project for the United States Department of Defense at Stanford University, Tableau is now an acclaimed desktop application which aspires to create complex but easy-to-read graphical interfaces for databases [5].

A visual analysis system, Tableau's functions are based on an algebraic specification language called VizQL which evolved from a system called Polaris, a system for query, analysis and visualization that was developed by two Stanford scientists [6]. Visual presentations of data are guided by a series of user-interface commands and defaults known as "Show Me." Thus, Show Me uses VizQL expressions to build visualizations.

Tableau accepts several types of tabular data ranging from Microsoft Access files to Microsoft Excel files to a text file of comma-delimited values. Once data is uploaded, data types are read by Tableau and are assigned to one of two fields and appear in the corresponding data windows: dimensions (typically categorical data) and measures (quantitative fields). Data categories appear as instances in each window.

Visual representations are created when users stipulate a VizQL expression by dragging an "instance" from the data windows to the column and row "shelves". Dimensions will appear as blue bars and measures as green bars. Once instances are "shelved," VizQL algebra clearly defines row and column structure to create visualizations of data.

Figure 2 -- Tableau visualizations

Tableau offers several visualization alternatives which include text tables, aligned bar graphs, stacked bar graphs, discrete lines, scatterplots, heat maps, scatter matrices, and Gantt

charts. To create a visualization, users click the button at the top of the screen labeled “Show Me”. In turn, Show Me will choose visualizations based on the conditions of the data and graph exponents from those considered to be “best practices”. At times, only one choice of visualization is offered to the user. Users may click an icon near the top of the screen, called “Show Me Alternatives” to gauge whether any other visualizations may apply.

Resulting visualizations appear as a tabbed workbook of sheets. Tableau allows users to save and export visualizations in Portable Document format (PDF).

The next section of this paper will describe the visualizations created for this study using these two tools.

III. Methodology

Many studies have been done regarding what makes a movie great – whether great in the eyes of the public or the academy or the box office. One of the more recent and closely related studies to this one was completed by Chris Weaver in his InfoViz 2007 contest submission [8]. In Weaver's study, he created an interactive visualization called Cinegraph using a visualization tool called Improvise. To create Cinegraph he set out to uncover answers to questions about the importance / impact of release dates, what genres have high box office hits, how Oscars related to genres, and if there are relationships between actors in movies and their box office and/or Oscar results.

Weaver's approach to uncovering these answers, or at least suggested answers, was able to be much more robust and thorough than this study. This was mainly due to the limitations of Many Eyes and Tableau as compared to Improvise. Even though similar kinds of data were used, Improvise allows for and almost seems to expect more relational data than a flat file data set, which is a definite benefit to Weaver's study. Even though Tableau can supposedly handle relational data sets, Many Eyes requires only flat file data sets and our team wanted to utilize both visualization tools for our research. Therefore, keeping this in mind, our team tried to focus on more statistical relationships, based off of totaled groupings, rather than individual results.

In order to do this, we first set out to define our data set. Utilizing a list of the Top 250 movies that was already voted upon and ranked by the IMDb community, we limited our set to only include the Top 100 movies. Using this as a starting point, we researched additional details about each movie to include the genre, Oscar nominations, Oscar wins, writers, directors, and country of origin as we felt that these items would potentially yield interesting results.

These additional fields accompanied the already provided ones from IMDb that included the title, release year, rank, and total votes. Additionally, we attempted to identify which Oscars were won versus which were only nominations, however, we later found that this would be too complex a level of detail for either Many Eyes or Tableau to handle. To acquire

the additional details we were after, each team member researched twenty-five movies. The research results were then combined into one file, formatted for consistency where necessary, and then loaded into Many Eyes and Tableau. Figure 3.3-1 shows an example of our data set format.

Rank	Rating	Title	Rel_Yr	Ttl_Votes	Genre	Director	Writer	O_Noms	O_Wins
1	9.1	The Godfather	1972	266645	Drama	Francis Ford Coppola	Mario Puzo	11	3
2	9.1	The Shawshank Redemption	1994	315395	Crime	Frank Darabont	Stephen King	7	0
3	9	The Godfather: Part II	1974	152715	Drama	Francis Ford Coppola	Mario Puzo	11	6
4	8.9	Buono, il brutto, il cattivo, Il	1966	87092	Drama	Sergio Leone	Sergio Leone	0	0
5	8.8	Pulp Fiction	1994	270702	Crime	Quentin Tarantino	Quentin Tarantino	7	1
6	8.8	Schindler's List	1993	181453	Biography	Steven Spielberg	Thomas Keneally	12	7
7	8.8	One Flew Over the Cuckoo's Nest	1975	135795	Drama	Milos Forman	Lawrence Hauben	9	5
8	8.8	Star Wars: Episode V - The Empire Strikes Back	1980	191380	Sci-Fi	Irvin Kershner	George Lucas	4	1
9	8.8	Casablanca	1942	118878	Drama	Michael Curtiz	Julius J. Epstein	8	3
10	8.8	Shichinin no samurai	1954	66525	Action	Akira Kurosawa	Akira Kurosawa	2	0

Figure 3 -- Sample Formatted Data Set

One additional data set was also used with the hopes of being able to confirm findings from IMDb or help us to understand differences. This data set included box office statistics from BoxOfficeMojo for the Top 50 highest grossing movies of all time. There was no need to manipulate this data to further enhance it because was to be used for comparison purposes only and was complete as possible without obtaining a paid subscription to the site. A minimal set of visualizations were done using this data, since as previously stated, it was used for overall comparisons only.

Our initial goal with this research was to find out what interesting relationships we could acquire amongst movie data user various visualizations. Specifically, based on our data set, we wanted to understand the following:

- what relationships exist between Oscar wins and nominations
- what genres of movies receive the most nominations versus what genres receive the most wins
- how nominations and wins relate back to public opinion
- what the public feels to be the best movies versus what the Academy thinks
- whether or not there are certain writers / directors that seem to fair better than others
- the possibility of the release year impacting the movie's overall rating by the public or the Academy
- how box office rates compare to overall and continued popularity of a movie

Of course, by referring to "the public", we are making a generalization based off of the IMDb user community voting results obtained for each movie.

Using Weaver's research tasks and questions as a base [8], and by understanding what data we would ultimately be able to work with after deciding on our research focus, we then set out to understand the available visualizations and how they would relate back to our data set. While we utilized a network diagram for some analysis, Weaver seemed to have much more success with them [8]. In addition to network diagrams however, we found

that bubble graphs, bar charts, and tree maps seemed to produce interesting and revealing visualizations given our data set. Many Eyes and Tableau describe the charts used as follows [5][6]:

- Network Diagram: identifies relationships between items and shows these relationships via linked nodes, with more strongly related items closer together and nodes with many links larger than those with fewer links.
- Bubble Chart: used to display data sets that have values with large differences between them so that bubble size can truly and more accurately relay the underlying differences with a strong visual impact.
- Bar Chart: best overall method to show generic comparisons within a data set and works best with discrete data sets.
- Tree Maps: allow for visualizations of hierarchical structures and can help show relationships amongst data where the structure is not immediately obvious, meaning that it require categories to exist naturally within the data that are uncovered by the visualization in order to be successful.

With our data set loaded into both Many Eyes and Tableau, and our goals identified, we began generating the various visualization diagrams to focus our analysis and interpret the results.

IV. Visualization Results

Many Eyes offered a variety of visualizations to display the data that was collected by the team. Through the data sets we were able to collect and generate numerous graphs which were created to show the linkage of the data between one another. Many Eyes requires the data being entered into the system to be publicly shown after publishing/generating the visualization. This system also requires the data to be in plain text form, not allowing for easy adjustment from Excel into Many Eyes.

Tableau works with data sets that are in Excel format, making the import of information easy on the end user. Tableau provides its own data sets that end users can chose to use instead of generating their own. Tableau gives the ability to import the files from Excel, keeping the transition of the data from one system to another as smooth as possible.

4.1 Top 100 Movies

We used IMDb, The Internet Movie Database, to provide us with a list of the top 100 movies over of all time. IMDb created a list of the top 250 movies, each with user rankings and ratings to determine the number one from the data. This data was in itself was revealing, showing what hundreds of thousands of users thought of the movies throughout time.

This data was placed into the Many Eyes program with the expectations of easy graphical viewing. Through the use of Many Eyes various errors were found. Though all of the users were able to upload data into the graphs that the system generated were not what were originally expected. Many Eyes does not warn against having

certain characters within the text files in order for the system to work properly. Java versions created no graphs to be shown and ‘Oops’ pages became more common than desired.

After the third attempt we were able to make our data work. From this data sheet we were able to create the bubble chart as shown in Figure 4. This chart gave a clear overview of the movies that received the highest number of nominations of Oscars and by their Ranking within IMDb. It allowed us to grab the data by the year they were made. There is no logic stated behind the placement of the bubbles within the graph. Years are scattered throughout the graph and the bubbles in the center are not higher ranked than the bubbles on the outer part of the circle. Many Eyes does not currently have logic behind the bubble placement, which gives a first skewed view of the results. At first glance, the viewer would focus on the center, making center seem more important than it may truly be.

Figure 4 – Bubble Diagram

Many Eyes offers various graphical representation options for data that is uploaded into the system. After the bubble charts were fully analyzed we reviewed other graphs, such as the Treemaps as seen in Figure 4.2. The Treemap chart was more appropriate for a basic understanding. The layout of the blocks was clearly defined within the top portion of the graphical over view. To adjust the block layout, the viewer needs to simply adjust the Treemap Hierarchy order to match their desired results. Many Eyes gives the options of genre, year released, director, writer, and title. It does not allow the viewer to block based off of rating, rankings and Oscar nominations. The color shading is listed at the bottom to give the viewer the option of how to adjust, allowing simple viewing for the end user. Here is where the viewer can adjust: the coloring for ranking, rating, total votes or Oscar nomination/wins. From this visualization it is plain that drama dominates as the genre for top movies by ranking. The drama category received the top Oscar nominations, and not surprisingly, the top Oscar wins. Fantasy, which had only one movie listed in this category, ‘Lord of the Rings: The Return of the King’, was the film with the most Oscar wins out of the top 100, with 11 wins for its final count.

Figure 5 – Treemap Diagram

The network diagram, shown in Figure 4.3, of Many Eyes allows the user to select what they want to link the movies together, and select the label they feel best suited. This graph is easy to spread out, adjust and focus in on. The vast variety of search and labeling

criteria allows for several different graphs, all showing the different aspects/relations between the top 100 movies. Affording this freedom gives the viewer the ability to section out their priority and gives a clearer visualization for their needs. Through the network view we were able to see the link of genre and top movies in an era. This allowed us to speculate that in each decade the genre favorites of the public change.

Figure 6 – Network Diagram

4.2 Box Office Statistics

IMDb offers various box office statistics for users to review. Within these statistics, there is the potential to show how a film's gross revenue can be affected by the genre that is released at a certain time of year. These trends were particularly interesting when laid out in a network diagram.

Tableau was used to pull past top movies box office data. This data consisted of all top grossing movies over the years and converting their earnings into present day currency. The graphical representation that we choose was the use a bar chart for this, as shown in Figure 4.4, believing this would give the clearest view of what the data was truly representing. We used the production companies of the films to break up the movies into certain sections. Top sellers from the top production companies were used to generate the graph, allowing all major production companies to show their top money makers. These findings, interestingly enough, did not coincide with IMDb's data of Top 100 Movies of all time.

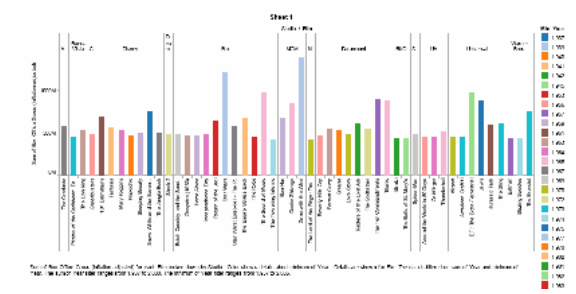

Figure 7 – Bar Chart

4.3 Comparison of Data

The top Oscar winner in the IMDb's top 100 movies was 'Lord of the Rings: The Return of the King', which was on the top grossing movies list. With under $500M, 'Lord of the Rings: The Return of the King' ranked far less than others in this group. The Tableau example showed the number one money maker of all time was 'Gone with the Wind', a movie not found in IMDb's Top 100 Movies. The reverse is also true. 'The Godfather' is the number one movie according to IMBD's ranking system, a movie not seen in the top grossing films. This shows a disconnect between each data set's information. Though both data sets have set formulas that validate their choice of number one films of all times, it is most interesting

that neither of the afore mentioned films even show up on the other data set all.

These data sets also showed how the genre of movie helped to increase the level of popularity. Using various visualizations we were able to clearly map out the hidden trends in the data. Genre popularity was affected by the year/decade the movie was released. Generating the graphs to be published by ranking and rating showed that Oscar nominations and wins did not affect the public's view of what was to be considered a good film. In truth, Oscars did not seem to affect films popularity whatsoever, nor did it seem to affect the amount of money generated by the overall films gross revenue.

V. Conclusion

The court of popular opinion is a difficult one to gauge. Every year, the film industry keeps meticulous detail of each film's box office earnings as an indicator of a film's success and popularity. Every year, the Academy of Motion Picture Arts and Sciences hold an extravagant awards ceremony where they award the crème de la crème of the previous year's films. Winning an Oscar is held by many as the highest achievement in film.

Still, it seems that these things have no consistent merit in a film's popularity with the public. 'Gone With the Wind' has the highest box office earnings of all time and won numerous Oscars, but does not appear on the IMDb's current Top 100 list. Meanwhile, 'Lord of the Rings: The Return of the King' won an amazing total of eleven Oscars and took in a staggering amount of money at the box office, and it does indeed appear on the IMDb list. The visualizations resulting from this data confirms the lack of reliability that box office earnings and Oscar attention contribute to cinema popularity.

The resulting data from this study, however, does indicate a few factors that seem to have an influence: release date and genre. 'Gone with the Wind' was released in 1939. So many decades later, its popularity may have faded a bit. 'Lord of the Rings: The Return of the King' was released only a few years ago. In another fifty years, who knows what this film's ranking will be. Genre also seems to play a part. Comedies are few and far between on the highest ranked films. They also tend to be overlooked by the Academy of Motion Picture Arts and Sciences. One consistent genre that the Academy and the public seem to agree on – Dramas, which rule in the top 100 listings.

Future studies may want to enhance on the direction of this study to better utilize the relational data capabilities of Tableau in order to dive deeper into what category of Oscar were won versus which were nominated for a particular movie. It may also be interesting to find more data on how IMDb users come to their rankings. Do they use a personal system or just a gut reaction? Such data for *why* a movie was ranked in IMDb that way could be standardized and used to enhance the visualization results to some extent.

Everyone has their favorite films and it does seem that there are a few common elements to the public's picks.

However, there is one main element that may prove possible to measure and visualize: individual taste.

References

[1] (n.a.). "IMDb History." Retrieved March 15, 2008, from http://www.imdb.com/help/show_leaf?history.

[2] (n.a.). (2007). "About Us: General." Retrieved March 15, 2008, from http://www.boxofficemojo.com/about/.

[3] (n.a.).(2001). "Collaborative User Experience." Retrieved March 15, 2008, from http://domino.watson.ibm.com/cambridge/research.nsf/pages/cue.html?Open.

[4] Collaborative User Experience Group. (2008). Many Eyes. http://services.alphaworks.ibm.com/manyeyes/home

[5] Tableau Software Group. (2007). Tableau Software. http://www.tableausoftware.com/

[6] Mackinlay, J., Hanrahan, P., & Stolte, C. (2007). "Show Me: Automatic Presentation for Visual Analysis." IEEE Transactions on Visualizations and Computer Graphics 13(6): 7.

[7] Viegas, F., Wattenberg, M., van Ham, F., Kriss, J. & McKeon, M. (2007). "Many Eyes: A Site for Visualization at Internet Scale." IEEE Transactions on Visualizations and Computer Graphics 13(6): 7.

[8] Weaver, C. (2007). InfoVis 2007 Contest Entry: Cinegraph. IEEE Visualization 2007 Conference Compendium, Sacramento, CA.