Analysis of Movie Rating Network

Fan Yang
fyang@cs.umd.edu
Department of Computer Science,
University of Maryland, College Park, MD 20742

November 8, 2012

1 Introduction

Movie archive websites such as IMDB provide users with a huge collection of movie resources including detailed information of movies and ratings given by numerous users. In this application project, I plan to visualize the network of movie rating and aim to find some interesting patterns which help us better understand the behaviors and preference of movie viewers.

2 Data & Tool

The data used is MovieLens 1M from http://www.grouplens.org/node/73 which consists of 1 million ratings from 6000 users on 4000 movies. Due to huge number of nodes and edges, I only sample a small portion of the entire dataset which includes 11398 ratings from 86 users on 2154 movies during a week in 2000. Each user rates a movie using integer score ranging from 1 to 5 where larger score indicates more positive evaluation on this movie, resulting a directed and weighted edge pointing from the user to the movie. All edges originate from users to movies so there are no relationship between movies and friendship between users.

The data formats are listed as follows.

- **Movie nodes.** A movie has three attributes: a unique ID, title and genre. There are 16 genres. To simplify the visualization, I only use the first genre label as the genre of a movie even though it may contain multiple genre labels.
- **User nodes.** A user has five attributes: a unique ID, gender, age, occupation and location zipcode. Gender is “M” or “F” representing male or female users. The age is categorical value indicating a predefined age group rather than numeric value. In my visualizations, occupation and location zipcode are not used.
- **Edges.** The edge is weighted by rating of a user on a movie. It also has an attribute indicating the time when the user rated the movie.

I use NodeXL to analyze the sample data and show the visualizations.
3 Insights

3.1 Young adults like to rate movies online than old adults

First, I group all movie nodes by the attribute “genre” and color different groups with different colors. The result is shown in the left subfigure in Figure 1, where size indicates the number of vertices in each group. It is clear that action, comedy, drama and horror are the most popular genres among users compared to that only a few ratings go to western, mystery and fantasy. However, it is hard to tell whether this situation comes from the preference of movie viewers or the overwhelming production numbers of such popular genres.

![Figure 1: Illustrations of groups of movie nodes and user nodes. Left: movie nodes are grouped by genre and size represent the number of nodes in each group. The groups are arranged in alphabetical order clockwise. Right: user nodes are grouped by age group and size indicates the number of nodes in each group. The blacks dots are movie nodes ordered by in-degree metric which measures the number of ratings received from users. In vertical direction, higher dots have larger in-degree value. Edges represent ratings from users to movies.](image)

Therefore, a deeper exploration is needed. Then I group all user nodes by the attribute “age”. After that, an interesting finding emerges that young adults from 25 to 34 are more likely to rate movies online than old adults, as shown in the right subfigure in Figure 1. The explanation is right at hand: young adults are more willing to use Internet to obtain information and express their thoughts while old adults are less likely to give comments through Internet. In 2000 when the data was collected, Internet was not so popular as it is today so old adults still depended on traditional media and could not use Internet skillfully.

Moreover, the genres of movies rated by 25-34 adults are more various than those rated by 56+ adults, which again indicates that young adults are willing to express their opinions and comment on movies through Internet. This finding reminds us that the online movie rating websites may be biased in ratings since young adults contributes far more than other age groups.

3.2 Age makes your movie taste change

The second finding just follows the previous finding. This time I only show two age groups: under 18 and over 56 and hide all the other groups. The two groups are two extremes among all age groups which represent the movie preference of juveniles and elders. I show the edges with rating score 5 which means that the user likes the movie very much.
In Figure 2, the two groups are shown as green and brown disks, and black diamonds in a line are movie nodes where darkness represents the number of nodes. Nodes are arranged using Sugiyama algorithms.

The pattern pops out right away: taste of movie viewers changes as they grow up. Although the two groups share interests in some movies, there is an obvious discrepancy between their other preferences (the left side and right side). After inspecting deeply, I find that users under 18 give more high ratings to movies of children, adventure, animation and comedy, while users over 56 love documentary, crime musical and romance. It is a natural conclusion since when you grow up you would not be interested in animation and children movies again.

3.3 The female movie addict tried more genres of movies than the male

In the last task, first I show all the vertices and edges in Figure 3. Triangles are user nodes and dots are movie nodes colored by genres. At first glance, it is hard to infer any useful information, so I focus on the users who are most active and contribute a large number of ratings. By this way, I aim to find some pattern of movie viewing and rating habits with respect to those “movie addicts” who are assumed to be eager and responsible in rating movies. I choose a male user and a female user who are coincidentally the top two users providing most ratings. The female user has rated 888 movies and the male user has a similar number 805.

In the left subfigure in Figure 4, we can see the two users share some movies, but they also have huge movie collections which do not overlap with each other. It means that male and female movie viewers indeed have different tastes although both have watched various genres of movies.

To better visualize the results, I further group movie nodes by attribute “genre” and rearrange the user nodes and movie groups (shown in the right subfigure in Figure 4) then reach the conclusion that female
movie addict dabbles in more various genres than male movie addict. More specifically, the female movie addict does not show a strong preference to some specific genres; she has watched all kinds of movies. In contrast, the male movie addict only selects a portion of movies; he does not rate any musical, romance, fantasy and mystery movies, etc., even though he has rated more than 800 movies. The most interesting thing beyond my expectation is that this female user has watched much more horror movies than the male user, which does not follow the usual case. This tells us our common sense is not always true.
4 Critiques

4.1 Good features

After using NodeXL, I find the following good features which greatly help my data exploration.

- NodeXL supports various input formats so I have a large number of datasets available to analyze.
- The Group by option is much helpful. Since my analysis is on a large network, doing grouping by different attributes provides me a few way to see the data from different perspectives.
- Autofill columns function and a number of different layout options also help me come up with good vertex arrangements. In particular, I like the functionality of defining edge width and opacity according to edge attributes.
- The automatic graph metric computation further facilitates my understanding on the structure of the network. By running computation and inspecting graph metrics, I quickly have a general sense of the characteristics of nodes and edges.

4.2 Bugs and things to be improved

There are also some bugs which result in unpleasant experience when I use NodeXL, listed as follows.

- First, the speed of redrawing graph, moving nodes and doing autofill is very slow when the network is large. This greatly affects the working efficiency and user experience.
- The layout speed using Sugiyama algorithm is significantly slower than other algorithms. Every time I make changes and layout nodes again using Sugiyama algorithm, I need to wait for a few minutes before I can see the results. If I am not satisfied with the layout, I need to wait for another few minutes. This is unpleasant, so a speed-up strategy is desirable.
- Sometimes I encounter errors when operating between autofill and refresh graph. A detailed exception list is shown but it is difficult for normal users to figure out what it means. A better exception handling could improve user experience when they get errors.