Influir – Artist's Influence Graph

Varad Pathak, Pranit Mhatre, Ujwala Tambe

University of Southern California
Computer Science Department
Los Angeles, CA 90089
{vspathak, pmhatre, utambe}@usc.edu

Abstract. Artists get influenced by each other. This is reflected in their artworks. Getting inspired by each others' work of art is common and indivisible to art world. The study of these influence graphs gives some insight into how the world of art works. Using the standard social networking algorithm we can calculate the level of influence of an artist based on the data about the influencing and influenced artists.

Based on this idea, we aim to construct an influence graph of filmmakers and provide a platform to explore a set of influential movies. This influence graph can be used to extract important information about the film industry and major influencing people in it. We have used this influence graph to find the movies which are most related to a given movie based on the influences on the director of the movie. This has resulted in a very good movie recommender system which finds the most influencing movie related to the given movie.

Keywords: Influence Graph, Movie, Director, Information Integration, InfluMeter.

1 Introduction

Influence graphs are commonly used in spreading or containing any information. It is used for marketing a product through influential people in the society and can also be used to contain contagious diseases or smoking habits. We have tried to apply the same principle to the influence graphs of artists. Artists get influenced by each other and add some of their own element to that and create new artwork which begets more influences. To generate more insight into this process of creation of artwork we are creating influence graphs of artists. As lot of related data is available about movies and influences of directors on each other, we have targeted to study the influence graph of film directors.

The influence graph of movie directors can help us find the most relevant movie which has influenced the director to create a given movie. So this way we can find a movie which has or may have a lot of influence on a given movie. We have used this idea to explore movies and thus have a movie recommender system. In the process we try to find directors who have influenced the given movie director and the directors who are most influenced by the given director.

Once we have all the related directors we need to narrow down the search using some additional clues like the genre of movie or keywords that describe the movie. This way we can predict the movie which has the highest influence on the given movie. So we rank them according to their level of influence on the given movie and we call this level of influence as InfluMeter Rating. Higher the InfluMeter Rating higher is the influence on the given movie.

2 Approach

We studied prior work on influence graphs and got better understanding different algorithms and approaches. Our work takes lot of clues from the work of Prof. Sinan Aral of New York University.

2.1 Algorithm

Professor Sinan Aral from NYU has done some important work on influence graphs constructed from social networks. As per his studies the "*Easily Susceptible users tend not to be influential, and influential tend to be stubborn*". This means that if a movie director who has influenced a lot of other directors and has been influenced by comparatively fewer directors is the most influential.

Other insight that we got from his studies is that there is a cycle of influences. "*Peer Adaptations beget Engagement and Engagement in turn begets more Peer Adaptations*". This means that directors get influenced and they create a movie which in turn influences more directors and cycle goes on.

These two insights are very important and are base of our application. From the first principle we deduce that the ratio of the 'is influencing' and 'is influenced by' directors is an important parameter of calculating an influence of a director.

$$M = \#\text{Influencing Directors}$$

$$N = \#\text{Influenced By Directors}$$

$$\text{Influence Factor} = M/N$$

This gives us an absolute influence of a director on the film industry. But we need a relative comparison of ratings of all the directors who have influenced / influenced by the given director. Thus we normalize the results by assuming that the most influential director among the given set of director has 100% influence score among all the directors and similarly we go on calculating the influence of a director on the given director.

Now we need to narrow down our results to find the most probable movie of an influencing director which has influenced the given movie. We achieve this by using additional data about the genre of the movie and the keywords that define the movie. So if there is more matching genre or keywords then there is higher chance that this particular movie of and influencing director has intrigued the given director and he added some of his own element to it to create a new movie.

Another factor which should be considered to calculate an influence of a movie (not director) is that 'what is its audience rating' and 'what its critic's rating is'. This is because among the given set of movies which are of same genre and of same director there is more probability that the movie with higher critics and audience rating will be more influential.

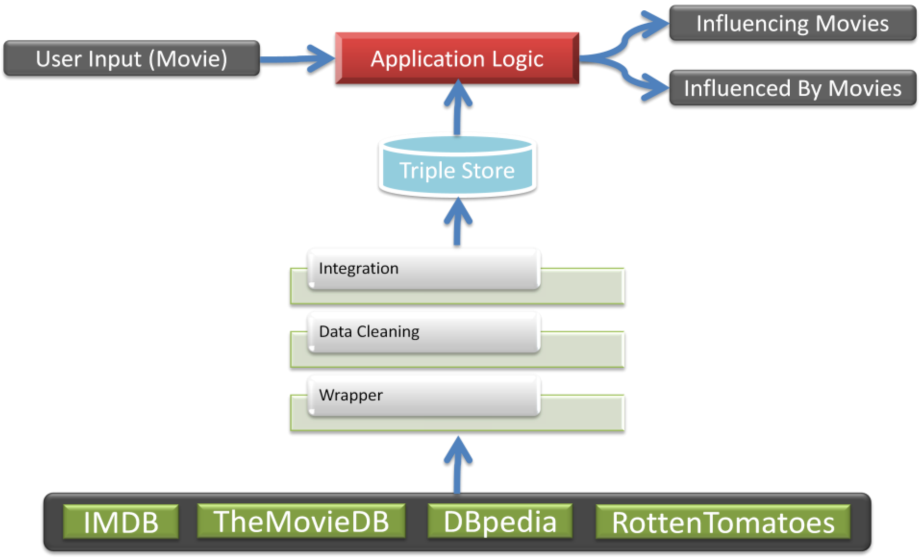
So our final formula to calculate an influence of a movie on another movie is a weighted average of following factors.

- Directors influence: 40%
- Genres: 15%
- Keywords: 5%
- Audience Rating: 20%
- Critics Rating: 20%

2.2 Architecture

The Fig.1 shows basic architecture of the system. As it can be seen the data is collected from different sources by creating a separate wrapper for each source.

Fig. 1.



Once the data is collected it is cleaned using Google Refine and the sanity of the data is verified as well. The cleaned data was integrated using the Fril tool for generating record linkages. In this step we used the Jaro Winkler similarity for string matching of the names of directors. We also use the Year of production of the movie to as an additional matching criterion. The integrated data is then modeled using Karma. The generated RDF from Karma is then uploaded to Triple Store at ISI.

We execute SPARQL queries on the triple store at runtime using the sesame api. This way we fetch the relevant part of the influence graph and then calculate the influMeter rating for each movie in the application. The results are shown using the Twitter Bootstrap API and JQuery library for UI.

2.3 Data Sources, Cleaning and Record Linkage

To limit the scope of our application we decided to create a complete influence network on top 250 movies (as per the IMDB ranking). So we collected the data about all the 250 movies from different sources like RottenTomatoes, TheMovieDB, IMDB using their corresponding APIs or file dumps. Then we constructed the graph of director influences up to 2 levels using the data from DBPedia using SPARQL queries. We also collected the data of all the movies which are directed by these newly added directors in the graph. We used the same set of sources for that.

Then we used Google Refine to clean all the collected data and resolved some of the issues like the movie name contain the year but the data about the year is missing. Few smaller issues had to be resolved in cleaning phase but mostly the data was clean, as the data sources used had very clearly defined APIs and almost every source follows the same standard.

Once all the data was collected we used Fril for record linkage of data that we got from different sources. The data collected from RottenTomatoes and TMDB had the IMDB IDs provided along with the other data through API. So that part of data integration was simple, we just compared the IDs. But the data collected from DBPedia which gives the details about the directors and their influences was not matching with the data collected from other sources. We used Jaro Winkler similarity for string matching the name of Directors and movies that we got from DBPedia and other sources. Along with this we used year of production as the other factor to make sure the movie referred in both the sources is exactly same.

Following are the details about the data that was collected from different sources.

2.4 Data Modeling

Once we got the clean data we constructed a mediated schema using Karma and modeled the data that we collected. We used some existing Ontologies to model our data but we needed some special classes and their relations, which were not defined by any of the available Ontologies, so we had to add some new classes and create their relations with existing classes.

We used following set of Ontologies to model out data.

1. dc: http://purl.org/dc/elements/1.1/
2. foaf: http://xmlns.com/foaf/0.1/
3. movieontology:http://www.movieontology.org/2009/10/01/movieontology.owl#
4. ontology: http://dbpedia.org/ontology/
5. dcterms: http://purl.org/dc/terms/
6. influir: http://influir.com/

3 Application Details

The application is developed majorly using java at the middle layer and JSP, CSS, JS in the frontend. During the data collection phase we used the RottenTomatoes API, TheMovieDB API and SPARQL queries on DBPedia. The collected data is always stored in JSON format to have the consistency and was helpful while integrating the data.

When the user queries details of a movie then the frontend gives the URI of the movie queried, this is then used to query the data stored in the Triple Store. The triple store is queried using SPARQL query which is executed using sesame api in a servlet. The results are then forwarded to frontend which displays the results using Twitter Bootstrap API and JQuery. The Twitter Bootstrap helps to have unified view across different devices and gives nice fluid look to the application.

Fig. 2. Landing Page, shows top 250 movies and direct search option

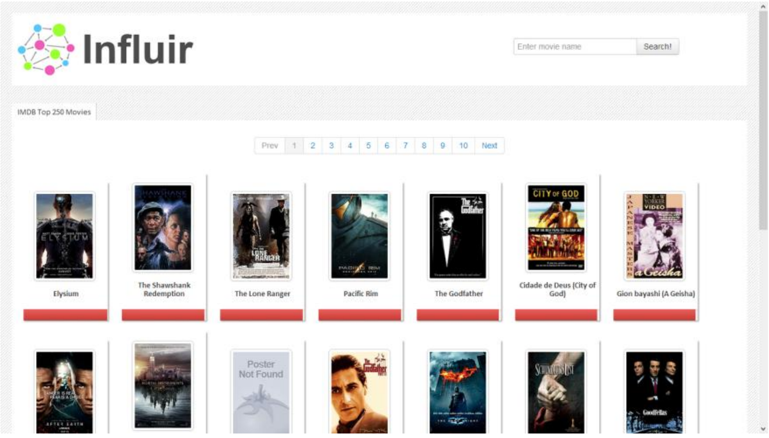


Fig. 3. Movie details page, Shows 'Influencing', 'Influenced By' Movies

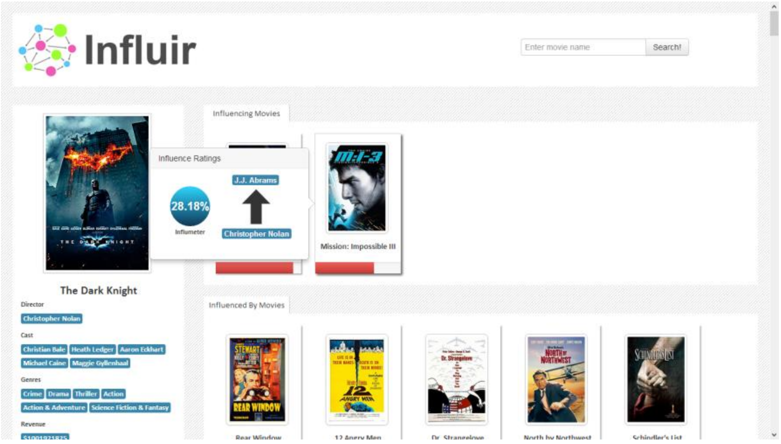
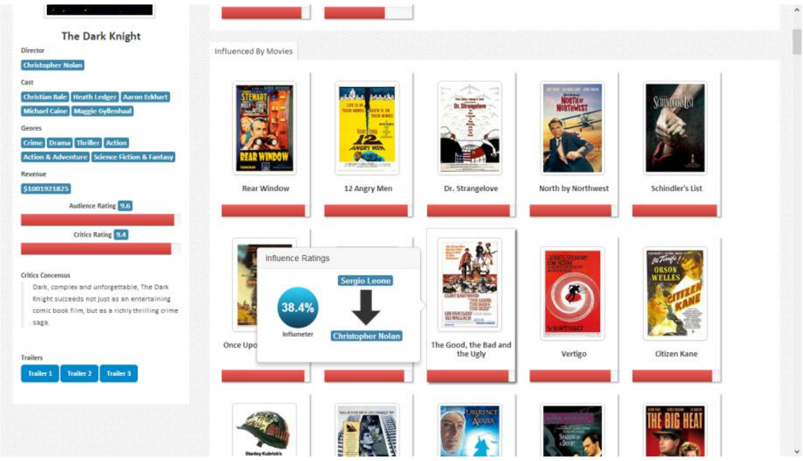


Fig. 4. Movie details screen, Shows other details and exact relation about the two movies through the popup and the influence rating of the selected Movie.



The Dark Knight

Director: Christopher Nolan

Cast: Christian Bale, Heath Ledger, Aaron Eckhart, Michael Caine, Maggie Gyllenhaal

Genres: Crime, Drama, Thriller, Action, Action & Adventure, Science Fiction & Fantasy

Revenue: $1001921825

Audience Rating 9.6

Critics Rating 9.4

Critics Concensus: Dark, complex and unforgettable, The Dark Knight succeeds not just as an entertaining comic book film, but as a richly thrilling crime saga.

Influence Ratings: 38.4% Influmeter; Sergio Leone → Christopher Nolan

The fig 2 shows the landing page through which user can search any movie among the top 250 movies and can also navigate through the pages. Once the user clicks any movie then the he moves to Screen 2 which is movie details page. This page shows the details about the movie such as the cast, director, revenue, audience and critic's ratings and critics' consensus about the movie and available trailers.

The right side of the Movie Details page shows influencing and 'Influenced By' Movies and if the user hovers his cursor on any of the movie then it shows its Influmeter Rating and exact relation among the two given movies that is the direction of influences among the two directors. If the user clicks on the movie then he can navigate to its movie details page.

4 Empirical Results

The underlying logic for computing and displaying the influences (as movies) on the screen is solely based on the value of its corresponding Influir score. In this section, using an example we will illustrate how this score is internally calculated. First, let us have a look at its computing formula:

$$\begin{aligned}\textit{Movie Score} = {} & ((0.4) * \text{influScore} / (\text{infMax} == 0\ ?\ 1 : \text{infMax})) \\ & + ((0.15) * (\text{genreScore} * 100 / \text{gSize})) \\ & + ((0.05) * (\text{keywordScore} * 100 / \text{kSize})) \\ & + ((0.2) * \text{criticScore}) + ((0.2) * \text{audienceScore})\end{aligned}$$

$$\textit{Influir Score \%} = \text{Movie Score}/100$$

Where,

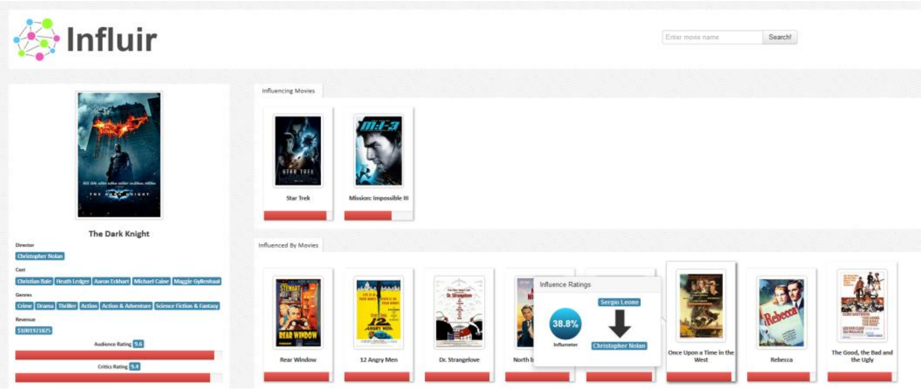
- *influScore* = ratio of the count of directors the candidate movie's director influences to the count of directors by whom the candidate movie's director is influenced by.
- *infMax* = maximum calculated InfluScore for a candidate movie. (In case of calculating influenced by movies, we get maximum InfluScore for an influenced by movie and vice versa for influenced by movie list).
- *genreScore* = Total match count of genre between the selected and a candidate movie.
- *gSize* = total genres of the selected movie
- *kSize* = total keywords of the selected movie
- *keyworkScore* = Total match count of keywords between the selected and a candidate movie.
- *criticScore* = critic rating
- *audienceScore* = audience rating.
- The floats in (0.x) are proportion of importance in calculating the overall InfluirScore.

(Note: Candidate movie is the one considered for influence/influence by movie for the selected movie.)

Based on this score we sort the movies in descending order to show them in the Influenced and Influencing movies grid.

Below is the illustration for 'The Dark Knight Movie'. When we click on the movie tile, the logic will first fetch all the movies which have its director to be influencing/ influenced by the director of the selected movie. Later based on the formula mentioned above, we calculate the Influir score and sort these movies for final display. The Influir score (for example) is calculated as shown below:

Fig. 5. Movie details screen, shows Influir score for The Good, the Bad and the Ugly



The Dark Knight

Director: Christopher Nolan

Cast: Christian Bale, Heath Ledger, Aaron Eckhart, Michael Caine, Maggie Gyllenhaal

Genres: Crime, Drama, Thriller, Action, Action & Adventure, Science Fiction & Fantasy

Revenue: $1001921825

Audience Rating 9.6

Critics Rating 9.4

Influence Ratings

38.8% Influmeter

Sergio Leone → Christopher Nolan

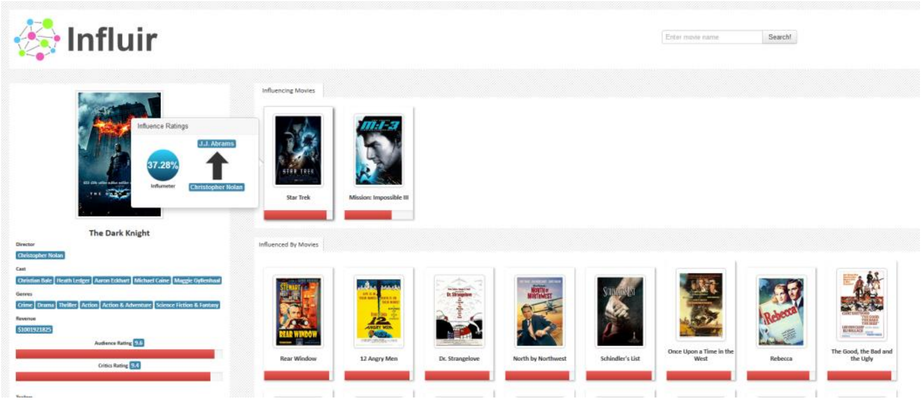
$$\textit{Influir Score} = (((0.4) * 5 / 5) + ((0.15) * (0 * 100 / 1))$$

$$+ ((0.05) * (0 * 100 / 16)) + ((0.2) * 97) + ((0.2) * 93))/100;$$

$$= 38.4$$

(Candidate movie: The Good, the Bad and the Ugly (influenced by movie), selected movie: The Dark Knight)

Fig. 6. Movie details screen, shows Influir score for Star Trek



The Dark Knight

Director: Christopher Nolan

Cast: Christian Bale, Heath Ledger, Aaron Eckhart, Michael Caine, Maggie Gyllenhaal

Genres: Crime, Drama, Thriller, Action, Action & Adventure, Science Fiction & Fantasy

Revenue: $1001921825

Audience Rating 9.6

Critics Rating 9.4

$$\textit{Influir score} = (((0.4) * 1 / 5)$$

$$+ ((0.15) * (0 * 100 / 1))$$

$$+ ((0.05) * (0 * 100 / 16))$$

$$+ ((0.2) * 95) + ((0.2) * 91))/100;$$

$$= 37.28$$

(Candidate movie: Star Trek (influencing movie), selected movie: The Dark Knight)

5 Related Work and Future Developments

There are several sites which give movie recommendations or show similar movies based on the plot of the movie and are mostly based on the manual entries. Examples are IMDB, RottenTomatoes, etc.

Also there has been work based on influence graphs in the marketing of products by few companies like Klout. Also some social networking websites like Facebook use the similar influence graphs to show ads based on influences.

In future we can collect some more data and perform analysis to get better results. We can movie references and details about the other connections like the sequel /prequel etc from IMDB to calculate the absolute influence of a movie and use it to refine the results.

6 Conclusion

We successfully constructed the influence graph of directors and were successful in getting the expected results. The Influence graphs can be used to get movie recommendations and get relevant results. If more data is provided such as the movie connections and references then the results can be improved.

7 Acknowledgements

The work on this project has been done as a part of class work for Information Integration on the Web (CSCI – 548) course at University of Southern California under the guidance of Professor Pedro Szekely and Professor Jose-Luis Ambite.

8 References

1. http://www.pnas.org/content/early/2009/12/09/0908800106.full.pdf
2. http://dbpedia.org/
3. www.rottentomatoes.com
4. www.themoviedb.org
5. www.imdb.com
6. http://dbpedia.org/sparql
7. http://www.dotnetrdf.org/api/~VDS.RDF.html