Knowledge Networks and Statistical Analysis of Cinematography Linked Data

Nikolaos Livanos
Open Knowledge Greece
Aristotle University of Thessaloniki
School of Mathematics
Thessaloniki, Greece
livanos@okfn.gr

Charalampos Bratsas
Open Knowledge Greece
Aristotle University of Thessaloniki
School of Mathematics
Thessaloniki, Greece
cbratsas@math.auth.gr

Sotirios Karampatakis
Open Knowledge Greece
Aristotle University of Thessaloniki
School of Mathematics
Thessaloniki, Greece
sokaramp@auth.gr

Ioannis Antoniou
Open Knowledge Greece
Aristotle University of Thessaloniki
School of Mathematics
Thessaloniki, Greece
iantonio@math.auth.gr

ABSTRACT

In this paper, we describe the methodology and the results of the statistical analysis, data mining, graph analysis and network analytics of cinematography data retrieved from the Linked Open Data cloud. The entire dataset was extracted with SPARQL queries, cleaned with Open Refine, and imported to R for data analysis. The methods applied were Cross Correlation, Mutual Information and Granger Causality. The results are visualized through an application that was created using R language's library "shinydashboard", giving the user the opportunity to view and edit the results with a set of widgets. This paper is a study that demonstrates the benefits of using Linked Open Data in Data Science and its methodology can be reused in many other multivariate Linked Open Data subjects.

CCS Concepts

•**Mathematics of computing** → **Time series analysis;** *Graph theory;* •**Information systems** → **Data mining; Data cleaning;** •**Applied computing** → **Mathematics and statistics;** •**Theory of computation** → *Data integration;* •**Computing methodologies** → Semantic networks;

Keywords

Linked Data, Knowledge Network, Statistics, Cinematography, Data Analysis

1. INTRODUCTION

The motion picture industry is one of the largest markets worldwide, grossing a total of 38.3 billion US dollars only in 2015[1]. For an amount of this magnitude, it is quite tempting to analyze data containing information about the financial situation of the distribution companies. However, obtaining such a large dataset can be a bottleneck for the average researcher. This paper serves as an example on how Linked Open Data can present a solution to this problem.

The data that was retrieved and analyzed contain information about published movies of the top movie distributors through the 15 year time period between 2000 and 2014. This dataset and its analysis is useful from many aspects, mainly for the distribution companies themselves, since through similar researches they can spot causes to their financial problems or make predictions amongst others [7].

Firstly, we describe step by step the procedure followed to retrieve and clean the dataset from the Linked Open Data cloud, along with the difficulties one will encounter conducting a similar work. Subsequently we explain the variety of tools used to analyze the retrieved data, along with the details of the application created to visualize the results. Finally, we comment on those results and offer insight on how this work can be a guide for many similar studies. The procedure followed is described on Figure 1.

2. DATA ACQUISITION AND CLEANING

Linked Open Data appear to be an ideal candidate for generating attributes to enhance statistical datasets, so that new hypotheses for interpreting statistics can be found [5]. Therefore, we choose the study the data for the top 16 film distribution companies, limiting by gross box office accounts. While LinkedMDB appeared to be the best source to retrieve this type of data, it still misses properties leading to valuable information, such as the budget and box office of every movie. Therefore, the entire dataset was retrieved using SPARQL queries to the endpoint of DBpedia [1].

Specifically, the properties acquired were:

- Movie Title

SEMANTICS 2016: Posters and Demos Track
September 13-14, 2016, Leipzig, Germany

[1] http://www.mpaa.org/wp-content/uploads/2016/04/2016-CinemaCon-.pdf

- Distribution Company
- Budget
- Box Office
- Year of Release

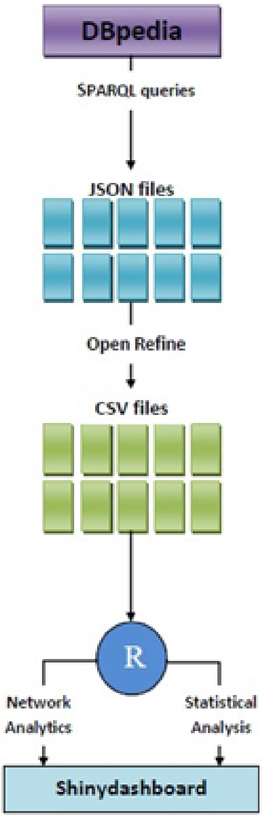

Figure 1: Procedure flow-chart

The retrieved dataset consists of 16 separate JSON formatted files, one for every distribution company, due to the limit of 10000 results per query returned from DBpedia's endpoint. These files were imported separately on Open Refine[8] for data cleaning. DBpedia data come from heuristic information extraction procedures, from a crowd-sourced web site, Wikipedia, leading to numerous types of errors[9].Many types of inconsistencies were found and corrected:

- Different wordings (Warner_Bros., Warner_Bros._Pictures)
- Numeric inconsistencies (10E6, 10million, 10000000)
- Duplicate entries
- Spelling Errors

For instance, the ”Year” column was edited using the ”text facet” option in a way that its cells only contain the 4 digits describing the year each movie was released. Both ”Budget” and ”Box Office” columns showed the same kind of anomalies and were edited similarly. In some cases the value was not described by US Dollars, and was converted for that reason manually using the existing equivalent. Even the ”Distribution Company” column contained many different wordings while referring to the same company. A unique wording was selected for every company, and the entire column was edited uniformly. The ”Movie Title” column was not edited mainly for the user's convenience to select any value and, this way, load the respective DBpedia page on his web browser. Finally, a new column was added, named ”Difference”, describing the profit (or loss for negative values) of each movie, calculated with the formula:

$$Difference = BoxOffice - Budget \tag{1}$$

Afterwards, the database was imported on R for analysis.

3. DATA ANALYSIS

We analyzed the retrieved and cleaned datasets over two different angles. The RDF data model implies a labeled directed graph by definition. So we examined the network structure of our graph. Then, we examined the dataset under statistical measures.

3.1 Network Analytics

Firstly, a network was created, using the distribution companies as vertices and the amount of their commonly distributed movies as the weight of the edges. For that reason, a 16x16 symmetric matrix was calculated. For every movie that appeared in the tables of two or more distribution companies, we added 1 to the respective elements of that matrix, as shown in Figure 2.

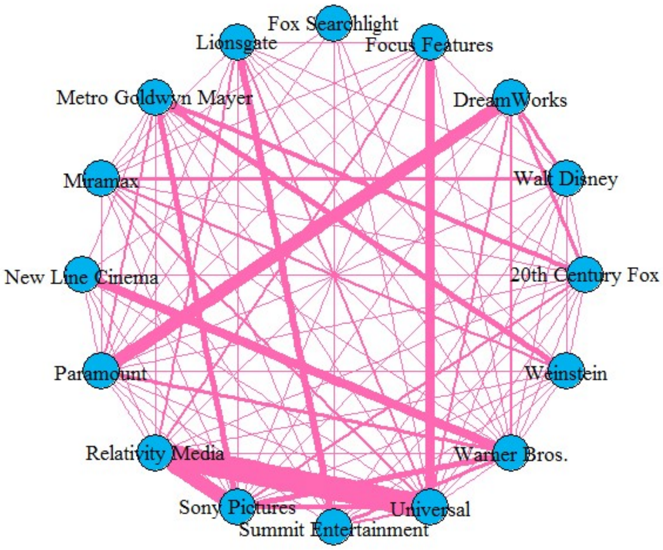

Figure 2: Amount of movies distributed by the same company

For a better understanding of the above network, we compared the sum of each row with the total number of movies every distribution company released in total. Consequently, a vector with 16 elements was calculated, showing the percentages of commonly distributed movies of the respective companies Figure 3.

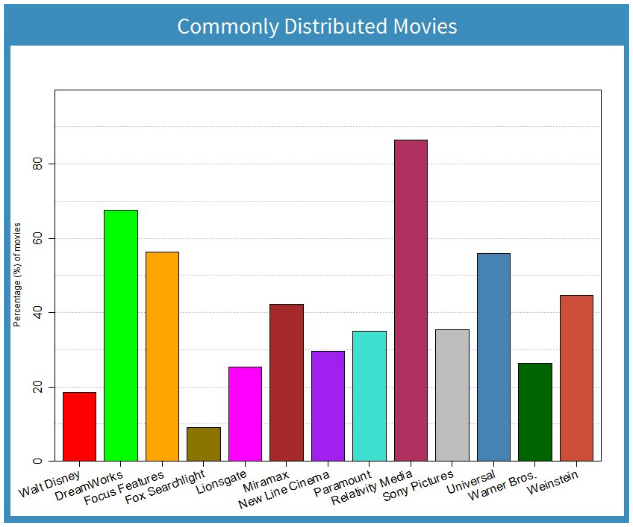

Figure 3: Percentage of movies distributed by two or more companies.

3.2 Statistical Analysis

For every distribution company, time series were created, using the total values per year of the "Difference" column, calculated before.

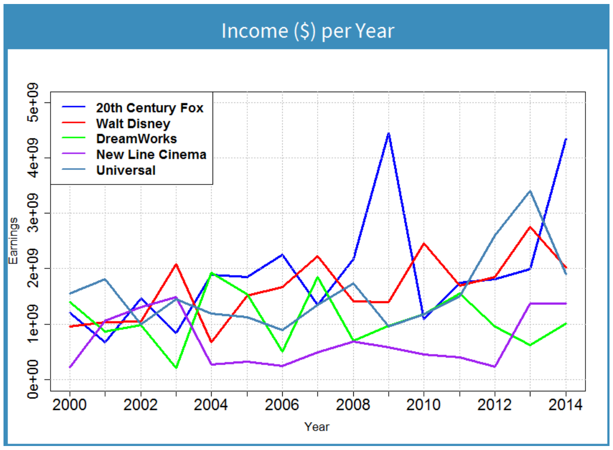

Figure 4: Time series visualized through the "shinydashboard" app.

The time series shown in Figure 4 were checked for correlations through

- Cross Correlation (Latency 0 and 1)
- Mutual Information [3]
- Granger Causality [4]

The cross correlation matrices for latency 0 and 1 were calculated, using the formulas:

$$r_{y_1 y_2}(k) = \frac{c_{y_1 y_2}(k)}{s_{y_1} s_{y_2}}; k = 0, \pm 1, \pm 2, ... \tag{2}$$

Where

$$c_{y_1 y_2}(k) = \begin{cases} \frac{1}{T} \sum_{t=1}^{T=k} (y_{1t} - \overline{y_1})(y_{2,t+k} - \overline{y_2}); k = 0, 1, 2, ... \\ \frac{1}{T} \sum_{t=1}^{T=k} (y_{2t} - \overline{y_2})(y_{1,t+k} - \overline{y_1}); k = 0, -1, -2, ... \end{cases} \tag{3}$$

Where $\overline{y_1}$ and $\overline{y_2}$ are the sample means of the series. The mutual information was calculated using the Local Gaussian Approximation.

$$\widehat{I}(X : Y) = \frac{1}{N} \sum_{i=1}^{N} log \frac{\widehat{f}(x_i y_i)}{\widehat{f}(x_i)\widehat{f}(y_i)} \tag{4}$$

Finally, the Granger causality was calculated for latency 1

$$X(t) = \sum_{T=1}^{L} A_T X(t-T) + c(t) \tag{5}$$

Every result was checked for its statistical significance:

- Cross correlation significance was checked through the p-value of its results. The p-value was computed by transforming the correlation to create a t statistic having 14 (n-2) degrees of freedom, where n is the number of rows of the time series matrix.
- Mutual information significance was checked through random iterations. Specifically 1000 iterations were made, and statistical significant values were considered the ones greater than the 95% (950) of them.
- Granger causality significance was also checked through the p-value for the F-tests computed to calculate the causality values.

4. VISUALIZATION

Clear and coherent visualization of Linked Data would enable broader accessibility to the Web of Data and encourage its use outside the SW community [2]. The results were visualized through an application created with R's library "shinydashboard"[2]. The application consists of 6 tabs in total: Network, Bar Chart, Pie Chart, Time Series, Statistics and Data. The first tab named "Network", displays a graph created using the matrix containing the amount of commonly distributed movies that was calculated before. The user can edit the vertices's and edge's size and color as well as the layout of the entire network. Moreover, a widget is included where the user can gradually erase the weaker edges of the network, allowing him to locate the strongest links easier.

The sum of the commonly distributed movies of each company, along with the respective percentages, are displayed in the "Bar Chart" and "Pie Chart" tabs.

In the "Bar Chart" tab, the user can select to plot a chart containing only the selected distribution companies.

In the "Pie Chart" tab, the user can select to see the percentage of the amount of commonly distributed movies for every company individually in a new pie chart with only two sectors. Every pie can be drawn in 3D as well for a more impressive visualization. The user can either spin or adjust the size of each chart. Specifically for 3D charts, the sectors can be "exploded" and the graph can be flipped across the third dimension.

[2] http://linkedanalytics.okfn.gr:3838/sample-apps/movies/

In the "Time Series" tab, the user can compare the total income per year of each distribution company through a line plot. More than one companies can be selected, and their respective lines will overlap in the same plot.

The "Statistics" tab, displays a graph with the distribution companies as vertices and the statistical values calculated from the time-series, see Section 3.2., as the weight of the edges. The user can select to view the graph of any of the three statistics included. Moreover, a check box is included where it's true value deletes every non-significant edges.

Finally, in the "Data" tab, the user can view, sort, search or download any part of the entire dataset he or she wants.

The source code along with the R workspace used to create the application are available in Github[3]

5. SUMMARY

The results showed that, to distribute a movie, the companies choose to collaborate with others far more often than expected. The average percentage of commonly distributed movies is 42.75%. That said, the respective variance is 456.7, meaning the values differ significantly from each other.

One would imagine that the above percentage would be more or less equal for every distribution company. As it turned out, that is actually far from true. The greater percentage of commonly distributed movies was Relativity Media's 86.39%. An interesting fact is that on July 30, 2015, the company filed for bankruptcy. Although many other factors could play their part on this situation, the value of that percentage definitely creates a case worth exploring.

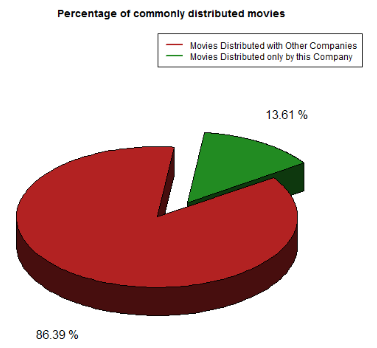

Figure 5: The case of Relativity Media.

The time series analysis, revealed many linear correlations between the distribution companies, many of which were statistically significant too. That applies for both latencies 0 and 1, but between different companies in each case. The mutual information showed a lesser amount of relations than the cross correlation, although not non existent. Finally, the Granger Causality showed the least amount of relations compared to the other two measures, meaning that one-way correlations between companies occur rarely.

6. FUTURE WORK

Many more studies can be conducted using the dataset retrieved in this paper. One can search for correlations between the distribution companies, using the most profitable collaborations as criterion, thus creating a strong advising tool for those companies.

On a different case, one can retrieve and analyze an entirely new dataset from the LOD cloud. Some cases include: music, literature, television and sports[6]. The distribution companies could be replaced with record companies, publishers, television channels and sports clubs respectively.

This paper can be used as a guide to highlight the difficulties one can encounter when using data retrieved from the LOD cloud. These problems can be faced with more targeted solutions this way, so that the time spent to conduct similar studies is minimized.

Acknowledgments

This work has been supported by the OpenBudgets.eu Horizon 2020 project (Grant Agreement 645833).

7. REFERENCES

[1] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. DBpedia - A crystallization point for the Web of Data. *Journal of Web Semantics*, 7(3):154–165, 2009.

[2] A. S. Dadzie and M. Rowe. Approaches to visualising Linked Data: A survey. In *Semantic Web*, volume 2, pages 89–124, 2011.

[3] S. Gao, G. V. Steeg, and A. Galstyan. Estimating mutual information by local gaussian approximation. *CoRR*, abs/1508.00536, 2015.

[4] C. W. J. Granger. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, 37(3):424–438, 1969.

[5] H. Paulheim. Generating possible interpretations for statistics from linked open data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7295 LNCS, pages 560–574, 2012.

[6] P. M. Philippides, C. Bratsas, A. Veglis, E. Chondrokostas, D. Tsigkari, and I. Antoniou. Creating and using sports linked data: Applications and analytics. In *CEUR Workshop Proceedings*, volume 1481, pages 38–41. SEMANTiCS 2015, Vienna, Austria, September 2015.

[7] S. Sparviero. The business strategy of hollywood's most powerful distributors: An empirical analysis. *Observatorio*, 7(4):45–62, 2013.

[8] M. Verlic. LODGrefine - LOD-enabled google refine in action. In *CEUR Workshop Proceedings*, volume 932, pages 31–37, 2012.

[9] D. Wienand and H. Paulheim. *The Semantic Web: Trends and Challenges: 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014. Proceedings*, chapter Detecting Incorrect Numerical Data in DBpedia, pages 504–518. Springer International Publishing, Cham, 2014.

[3] https://github.com/NLivanos/Cinema-Linked-Data