BMKG Project Report

Movie Recommendation system built on a integrated library of OTTs

Advaith Jaishankar(i6274688)

March 30, 2022

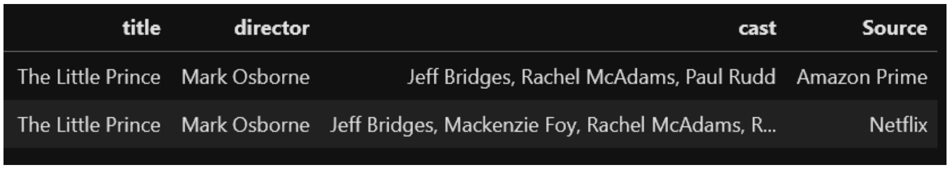
1 Abstract

The rise of OTT platforms and the presence of a non overlapping libraries across them has led to a need for an integrated recommender system based on knowledge graphs which is what this project aims to achieve. A knowledge graph linking the movies on OTT platforms and the ratings given by users for the various titles was built and a recommender system based on the link prediction method was built on top of this knowledge graph. Two models were tried out and it was discovered that the TransE model outperformed the ComplEx model. For future work, it is possible to extend this to other media like TV Series or even music provided that we have a dataset containing the ratings that each title received.Also other approaches to build recommender systems can be tried.

2 Significance

A recent trend has been the rising popularity of online streaming services also known as Over-the-top(OTT) media services as an alternative to the more traditional cable and satellite television. Popular examples include Netflix,Prime Video and Disney+. The various OTT platforms often have content which is exclusive to the platform making it difficult to consolidate the information present across the various platforms.

The rise of OTT platforms is mostly driven by the options available on the platforms and the customer service offered by these platforms[LNRL18]. Also the penetration of mobile internet and smartphones which support the viewing of online content alongside the production of localised content are cited as causes for the rising popularity of these services[SN20]. These streaming services often have content that is unique to the platform with little overlap to the other platforms. Additionally even for the same movie, the information present can be inconsistent. For example, an actor might be uncredited in a movie present on Prime but might be credited on Netflix. An example of such an inconsistency is shown in the figure 1. Here we can see that for the same movie, only 3 actors are credited on Amazon Prime but a lot more actors are credited on Netflix. The creation of a knowledge graph aims to resolve such inconsistencies. These factors make it difficult for the consumer to access consolidated information about the movies present on the other platforms easily.



title	director	cast	Source
The Little Prince	Mark Osborne	Jeff Bridges, Rachel McAdams, Paul Rudd	Amazon Prime
The Little Prince	Mark Osborne	Jeff Bridges, Mackenzie Foy, Rachel McAdams, R...	Netflix

Figure 1: Inconsistencies in the data

Recommendation engines are important to the OTT because it provides useful and relevant recommendations to the user thereby providing a personalized experience to the user and keeping them engaged for longer. They also help build a consistent brand experience where the customer knows what to expect when he uses the platform since it also stores their history rather than it feeling like a fresh experience every time. These factors together help build brand loyalty among the customers and help the platforms retain their customer base and even increase it.

This project proposes the creation of a knowledge graph integrating the data available about the libraries of Netflix[Net22], Prime Video[Pri22], Hulu[Hul22] and Disney+[Dis22]. Additionally, I aim to build a recommender system using the MovieLens[Mov22] dataset by linking the knowledge graph built from that dataset with the one built previously.

3 Related Work

This project builds upon some previous concepts introduced by other authors. Namely the knowledge graph of movies has been implemented by [HC09]. Some of the techniques like approximate join techniques will be used in this project as well. What differentiates this implementation from that of the paper is that there is focus solely on integrating the movies streamed across OTT platforms and linking them to other ontologies but the author of this paper uses information from other open data sources like FreeBase and OMDB. This however has not been updated for a couple of years.

The recommender systems that will be used in this project have been previously proposed by other authors as well. A comparison of the different possible approaches has been explored in [GZQ+20]. In particular, the approach using graph embeddings to create a recommender system has been explored in [AACZ18] and an approach based on the relations between the different entities has been explored in [SZL+15].

4 Goal and specific objectives

The main goal of this project is to create a knowledge graph from the various datasets regarding the content streamed on the platforms and integrate this knowledge graph with another knowledge graph created from the MovieLens dataset and derive insights from the integrated knowledge graph. The intention behind linking the knowledge graphs is to create a recommender system which recommends movies to users based on the other movies that they have previously liked. The main specific objectives of the project are:

- Create a knowledge graph by integrating the datasets of the content available on the various OTT platforms
- Create a knowledge graph from the MovieLens dataset which has the title of the movie, the various users that rated it and their ratings
- Link the two knowledge graphs so that the knowledge is shared across the two knowledge graphs
- Implement a recommendation system on the linked knowledge graph to output movie recommendations based on user history

5 Methodology

First, we concatenate the datasets of Netlfix, Hulu, Disney+ and Prime Video found on Kaggle and create a knowledge graph from it containing information like the streaming platform, cast members and director. The dataset obtained contains both movies and TV Series but to delimit the scope of this project, only movies have been taken into consideration and TV Series were not included in the knowledge graph. This is because the MovieLens dataset has information pertaining to only movies. So including TV Series would not add any value. Next I will create a knowledge graph using the MovieLens. The version of the dataset chosen is the *Small* version which contains around 9,000 movie titles and 100,000 ratings given by 600 users. This knowledge graph will contain the movie titles and will link users and their ratings to it. For this project,the reviews of only the movies that are present on OTT platforms are used.

The next step is to link these two knowledge graphs using the title of the movie as a common link. The aim of this to allow the transfer of knowledge like the cast members and director of a movie to the knowledge graph which contains the ratings of the movie. The actors present in the movie and the director may influence a person's decision to watch a movie or not and this is a crucial factor for the recommender system that will be built next. This was done using fuzzy string matching by comparing the names of the titles in both datasets and using the year of release as another parameter to avoid any mismatches. Ontologies like Schema have been used to define the types and Example has been used to create some custom relations between the entities.

Owing to the lack of time and other constraints, only one approach was implemented and evaluated. Link prediction was used. Both ComplEx[TWR+16] and TransE[LLW+17] models have been implemented and the hyperparameters used are in accordance with the optimal hyperparameters as mention in the documentation of Ampligraph[CPV+19] which is the library used to work with knowledge embeddings in this project.

6 Milestones and Deliverables

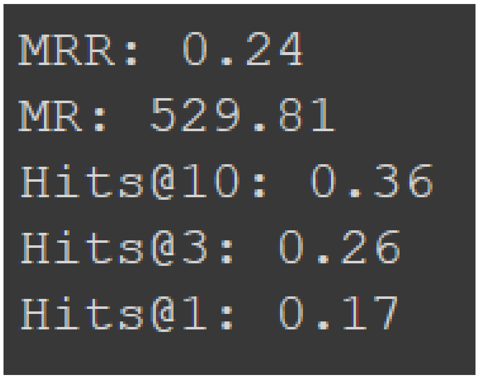
1. **Knowledge graphs constructed and linked (Week 2).** Both the knowledge graphs have been formed and linked using the concepts mentioned above.

2. **Recommender system complete (Week 3).** The recommender system is complete and an approach has been tried and metrics reported.

3. **Documentation (Week 4).** The documentation along with the code is uploaded to the Canvas portal in a zip file.

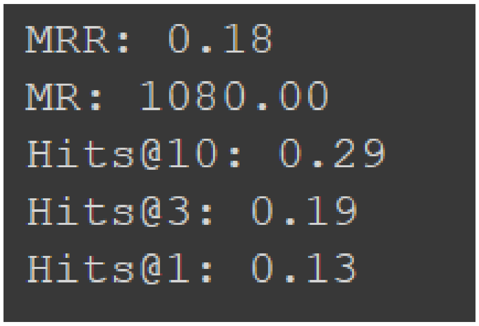
7 Results

The knowledge graph was built by linking the two datasets and a sample of the linked knowledge graph for just one movie is attached with the project deliverables. The link prediction method along with two different models were implemented and compared using the metrics - Mean Reciprocal Rank(MRR), Mean Rank(MR) and Hits@N scores and the results are shown below. As we can see from the images2 and 3,the TransE model has outperformed the ComplEx model. Also an integrated knowledge graph was produced which is attached in the project deliverables as a ttl files and the ipynb notebooks which were used to process the data and create the models are also included in the project deliverables.



```
MRR: 0.24
MR: 529.81
Hits@10: 0.36
Hits@3: 0.26
Hits@1: 0.17
```

Figure 2: Results for TransE model



```
MRR: 0.18
MR: 1080.00
Hits@10: 0.29
Hits@3: 0.19
Hits@1: 0.13
```

Figure 3: Results for ComplEx model

8 Discussion

8.1 Discussion of Results

As seen in the above section, the TransE model performs better and has better predictive power as indicated by the metrics - Mean Reciprocal Rank(MRR), Mean Rank(MR) and Hits@N scores which is counter intuitive as ComplEx model would be expected to perform better.

8.2 Challenges

Since the names of the titles were different in both datasets, fuzzy string matching had to be done in order to match the names so that they can be combined to create a single dataset. Because of this process, a large number of movies which did not have reviews for them had to be eliminated while forming the final dataset. Owing to the time constraint, two approaches were not implemented as discussed in the proposal.

8.3 Future Work

In the future, the same work can be extended to include more movies and even TV Series provided that the dataset containing the ratings for them is available. Also other models and approaches can be implemented which were omitted in this project because of the time constraints.

9 Conclusion

This project's main goal was to expand the body of knowledge available by linking the two datasets considered and creating knowledge graphs from them. Additionally, a recommender system was built on top of the said knowledge graph using the link prediction approach and it was discovered that among the two approaches used, ComplEx was outperformed by TransE model. The scope of this project need not be limited to just movies but can also be expanded to include TV Series provided that we have a dataset that contains users and their ratings of the content.

References

[AACZ18] Qingyao Ai, Vahid Azizi, Xu Chen, and Yongfeng Zhang. Learning heterogeneous knowledge base embeddings for explainable recommendation. *Algorithms*, 11(9), 2018.

[CPV$^+$19] Luca Costabello, Sumit Pai, Chan Le Van, Rory McGrath, Nicholas McCarthy, and Pedro Tabacof. AmpliGraph: a Library for Representation Learning on Knowledge Graphs, March 2019.

[Dis22] Disney+. https://www.kaggle.com/shivamb/disney-movies-and-tv-shows, March 2022.

[GZQ+20] Qingyu Guo, Fuzhen Zhuang, Chuan Qin, Hengshu Zhu, Xing Xie, Hui Xiong, and Qing He. A survey on knowledge graph-based recommender systems, 2020.

[HC09] Oktie Hassanzadeh and Mariano P. Consens. Linked movie data base. In *LDOW*, 2009.

[Hul22] Hulu. https://www.kaggle.com/shivamb/hulu-movies-and-tv-shows, March 2022.

[LLW+17] Hailun Lin, Yong Liu, Weiping Wang, Yinliang Yue, and Zheng Lin. Learning entity and relation embeddings for knowledge resolution. *Procedia Computer Science*, 108:345–354, 2017. International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland.

[LNRL18] C. Lee, Pankaj Nagpal, Sinead Ruane, and Hyoun Lim. Factors affecting online streaming subscriptions. 16:Article 1, 10 2018.

[Mov22] Movielens. https://grouplens.org/datasets/movielens/, March 2022.

[Net22] Netflix. https://www.kaggle.com/shivamb/netflix-shows, March 2022.

[Pri22] Amazon prime. https://www.kaggle.com/shivamb/amazon-prime-movies-and-tv-shows, March 2022.

[SN20] E. Sundaravel and Elangovan N. Emergence and future of over-the-top (ott) video services in india: an analytical research. *International Journal of Business Management and Social Research*, 8:489–499, 01 2020.

[SZL+15] Chuan Shi, Zhiqiang Zhang, Ping Luo, Philip Yu, Yading Yue, and Bin Wu. Semantic path based personalized recommendation on weighted heterogeneous information networks. pages 453–462, 10 2015.

[TWR+16] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction, 2016.