URBAN/ECO Research Centre
University of Naples “Federico II”

University of Catania

A Multi-source graph representation of the Movie domain for Recommendation Dialogue Analysis

Antonio Origlia, Martina Di Bratto, Maria Di Maro, Sabrina Mennella
{antonio.origlia, martina.dibratto, maria.dimaro2}@unina.it, sabrina.mennella@phd.unict.it

Introduction

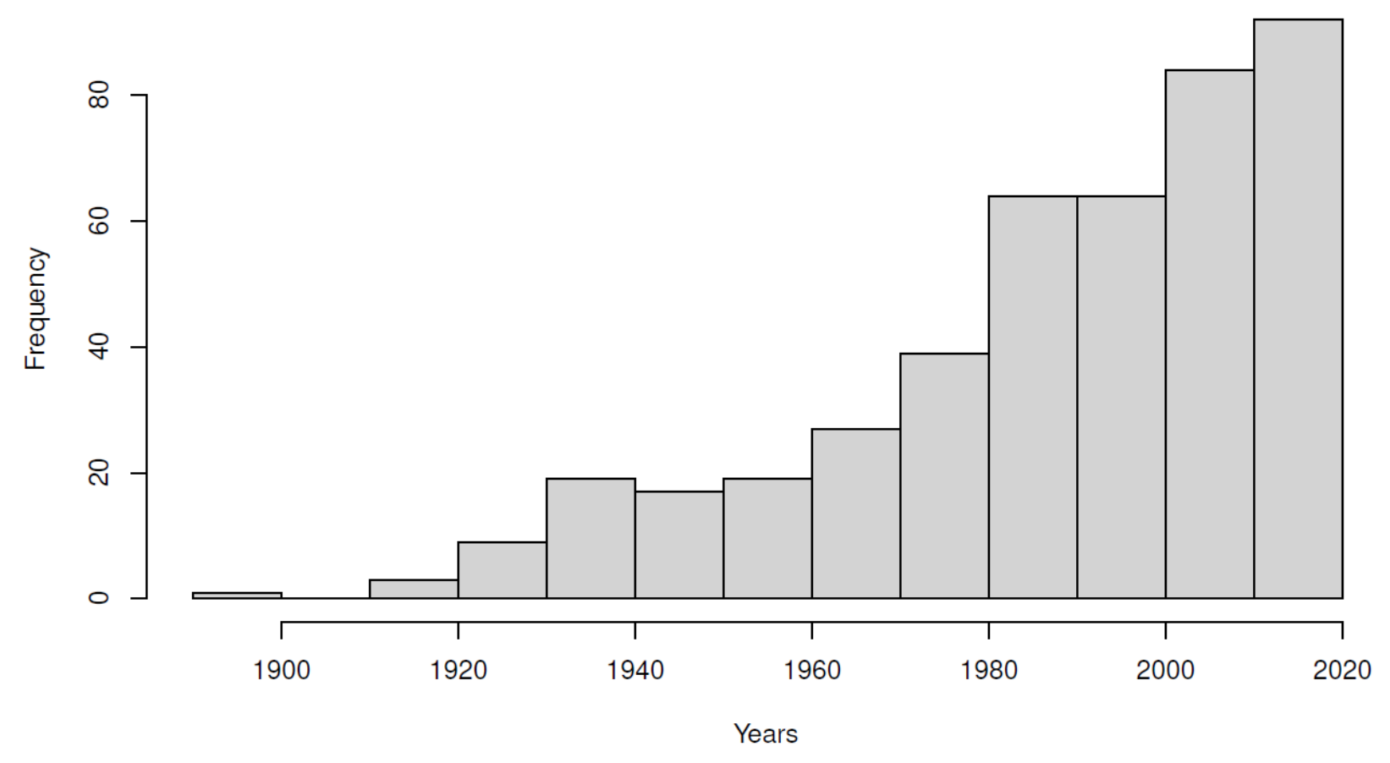
Graph databases have gained popularity for their capability to efficiently represent complex data in an interpretable, flexible format. Graph databases are well-suited in cross-referencing different resources so that information distributed in multiple sources can be analysed in an integrated way. Graph databases have found application in different fields requiring the management of large datasets characterised by complex interactions among items.

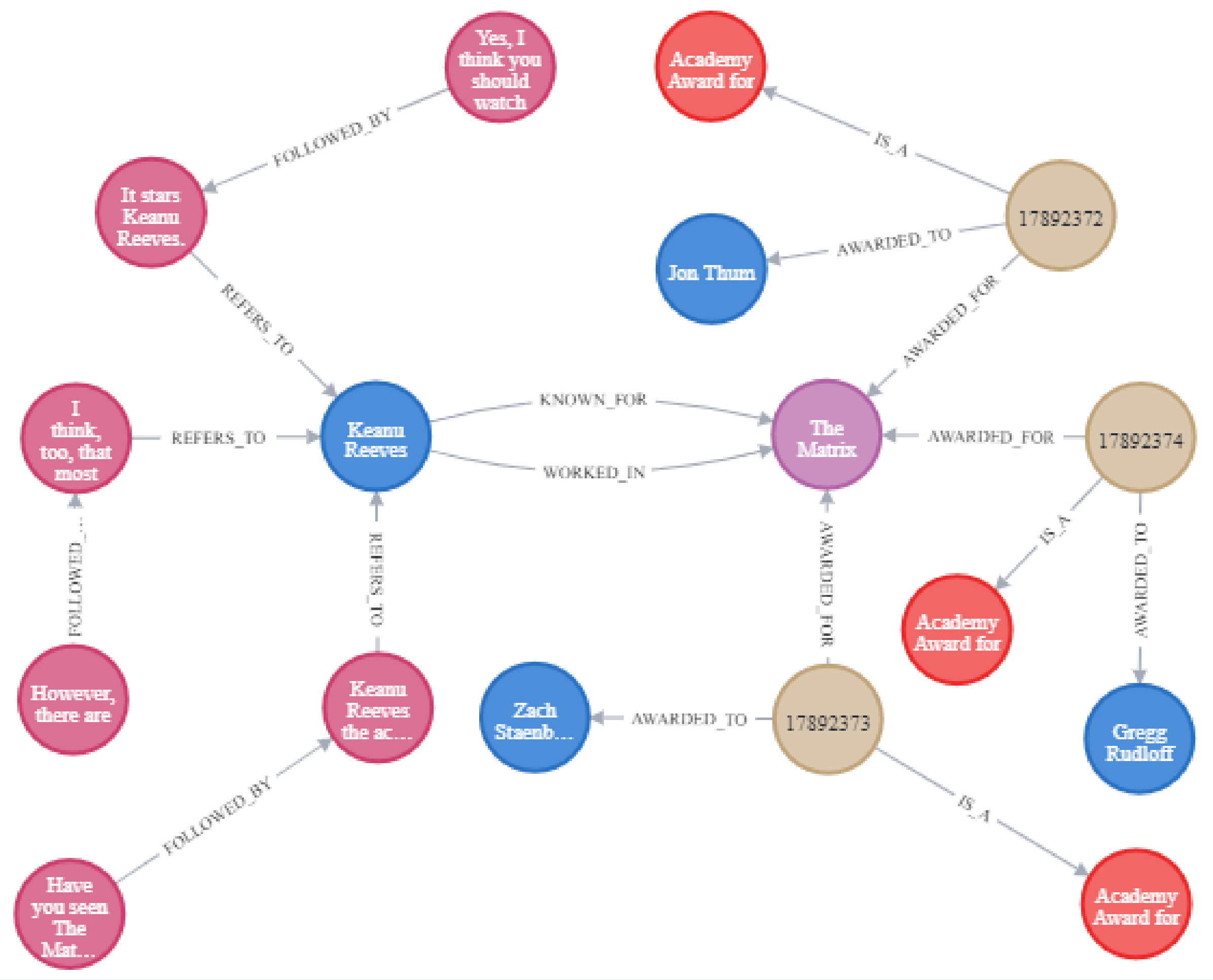
The movies domain represents a well studied case of interest for a number of different tasks.

As there are many different approaches about this specific domain, we present a data processing pipeline to integrate, in a single, graph-based, resource, the information contained in some of the most relevant resources that are available online. Also, we apply graph analysis techniques to enrich the obtained network with latent information extracted from the final graph structure.

The graph structure

An extract of the obtained structure after data ingestion, using the Neo4j database (Webber, 2012). The different sources are organised in a single graph structure optimised for graph traversal queries, to cross-reference information that would otherwise be distributed in different resources.





Materials

The Internet Movie Database (IMDb) is the most popular dabatase for movie, TV, and celebrity contents. It includes more than 8 million films and programmes, and about 10 millions actors.

IMDb does not provide an ontological view of movies and it does not report information about awards won by people and movies. Such information is found in Wikidata, which also contains the IMDb ids for both movies and persons.

The Inspired corpus (Hayati et al., 2020) is a dataset containing 1001 recommendation dialogues of two-paired crowd-workers who chat in a natural setting in English.

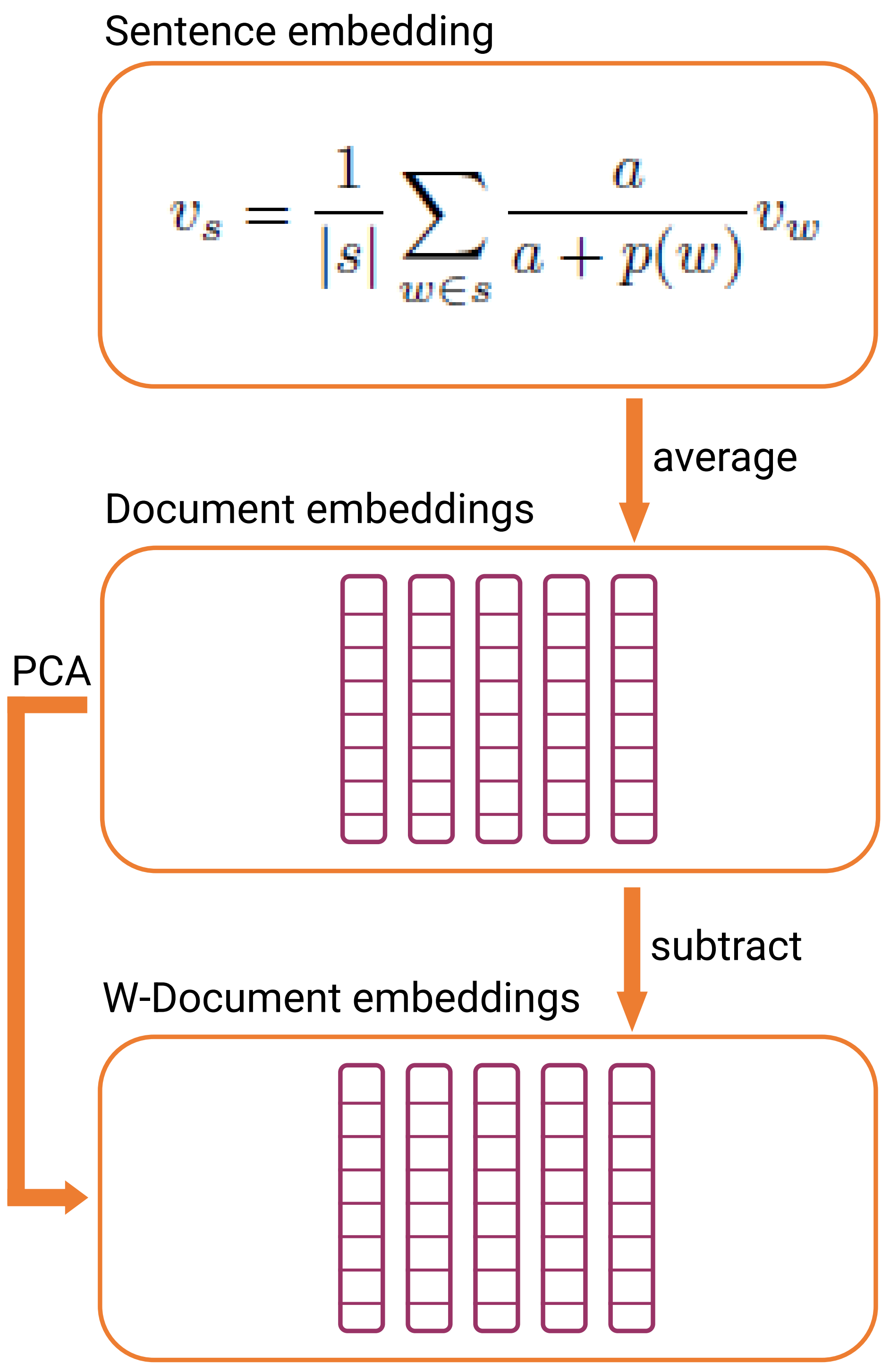
User evaluations about movies are taken from the Movielens 25M dataset (Harper and Konstan, 2015). This contains 25000095 ratings and 1093360 tags across 62423 movies.

Enrichment

To compute textual similarities in the Wikipedia articles, we consider the results shown in (Ranashinghe et al., 2019), which demonstrated that, on the sentence similarity task, a stacked embedding composed of ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018) embeddings outperformed other solutions when results were weighted by Smooth Inverse Frequency (SIF) (Arora et al., 2017).

PageRank is also computed to support disambiguation in identifying the subjects of Inspired utterances, especially in the case of homonyms.

Lastly, node embeddings are computed using the Fast Random Projection algorithm to summarise the semantic role of the graph nodes.


Sentence embedding

$$v_s = \frac{1}{|s|} \sum_{w \in s} \frac{a}{a + p(w)} v_w$$

average

Document embeddings

PCA

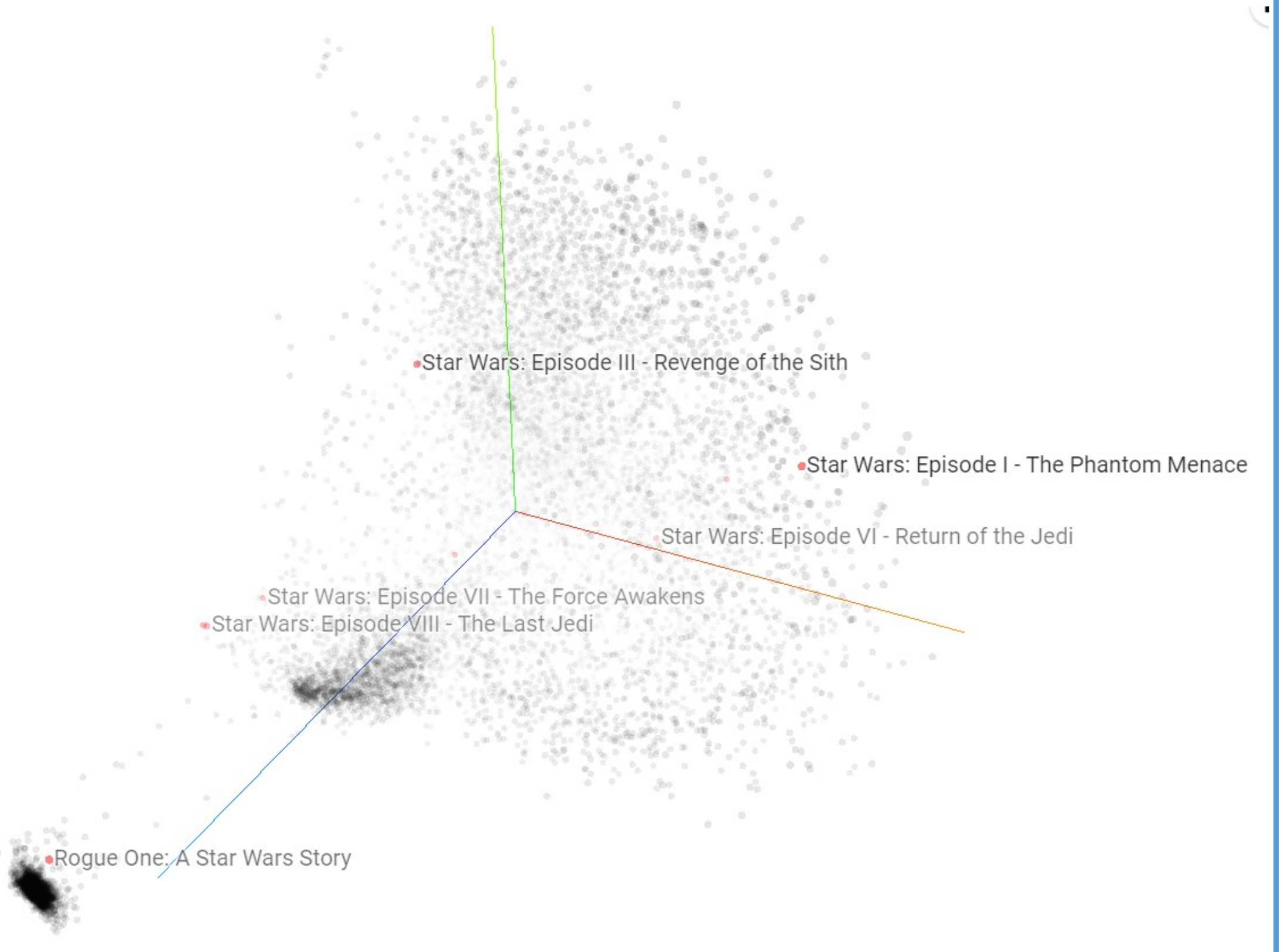
subtract

W-Document embeddings

Results

Considering the case of the Star Wars saga, the represented movies appear to occupy the same space, on the basis of catalographic information only. When considering unweighted SIMILAR_TO relationships, the same movies appear to group together following the different trilogies.

The database supports the study of argumentation based dialogue while considering technical aspects concerned with the dialogue systems design.



Conclusions

We presented the assembling process of a graph database combining multiple data sources into a single structure. The database cross-references a corpus of movie recommendation dialogues with domain knowledge collected from different datasets. We also described a data enrichment procedure using text processing techniques and graph data science algorithms to encode complex information into compact representations that can be later used for machine learning tasks. While we cannot distribute the database, as it includes data coming from other sources, we provide the source code to assemble it once the appropriate licences have been obtained. We also provide the results of the data enrichment procedures so they can be directly imported in the database without recomputing the results, some of which required significant computational power.

References

[Arora et al., 2017] Arora, S., Liang, Y., Ma, T.: A simple but tough-to-beat baseline for sentence embeddings. In: Proceedings of the 5th International Conference on Learning Representations (2017)

[Devlin et al., 2018] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

[Harper and Konstan, 2015] Harper, F. M. and Konstan, J. A. (2015). The movielens datasets: History and context. Acm transactions on interactive intelligent systems (tiis), 5 (4):1–19.

[Hayati et al., 2020] Hayati, S. A., Kang, D., Zhu, Q., Shi, W., and Yu, Z. (2020). Inspired: Toward sociable recommendation dialog systems. arXiv preprint arXiv:2009.14306.

[Peters et al., 2018] Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. arXiv preprint arXiv:1802.05365 (2018)

[Ranashinghe et al., 2019] Ranashinghe, T., Orasan, C., Mitkov, R.: Enhancing unsupervised sentence similarity methods with deep contextualised word representations. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (2019)

[Webber, 2012] Webber, J.: A programmatic introduction to neo4j. In: Proceedings of the 3rd annual conference on Systems, programming, and applications: software for humanity. pp. 217–218. ACM (2012)