A hybrid filtering movie recommendation system

Yuntian He, Zhengqi Dong

Dec. 4th, 2020

Dataset description

- **Movies** from Kaggle
  - **26M** ratings from **270K** users on **45K** movies
- Content
  - **Text**: Each movie has an overview (a paragraph) and some tags
  - **Rating**: A tuple (User, Movie, Rating, Timestamp)
  - **Attributes**: Each movie has multiple attributes including
    - Genre
    - Credits (cast/crew)

* https://www.kaggle.com/rounakbanik/the-movies-dataset

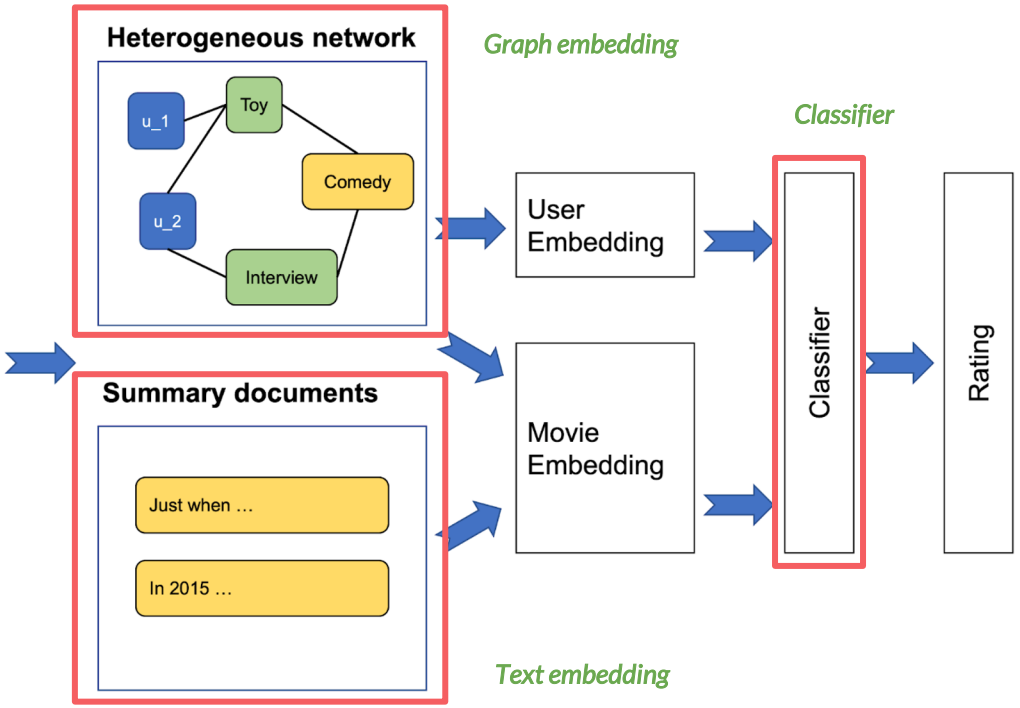
System framework

Ratings

User	Movie	Rating
u_1	Toy	4.5
u_2	Toy	3.0
u_2	Interview	5.0
…	…	…

Movies' metadata

Movie	Genre	Summary
Toy	Comedy	Just when…
Interview	Comedy	In 2015, …
…	…	…



Preprocessing

- Removing data in incorrect format
  - **3** of 45K movies are deleted
- Index adjustment
  - Consecutive IDs for convenience
- Attribute selection
  - **Cast**: Only top 8 casts (cast order included in the raw data)
  - **Crew**: Only use ‘director’

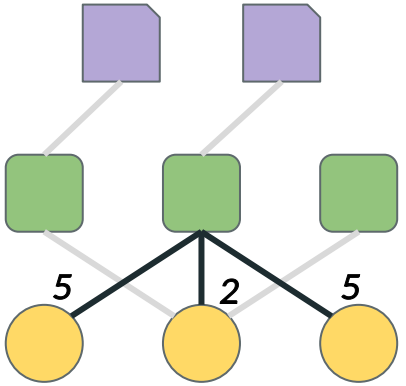
Text embedding

- Doc2vec
- BERT
- …

- Like a black box

Graph embedding: Metapath2vec

- Heterogeneous information network
  - User (**U**), Movie (**M**), Genre (**G**), Cast/crew (C)
- Metapath-based sampling
  - Preserve semantic relationships between nodes
  - **U-M-U**, **U-M-G-M-U**, **U-M-C-M-U**
- **Rating-aware** sampling policy



Genre

Movie

User

$$P(s_{t+1} = m | s_t = u) = \begin{cases} 1/|N_M(u)| & , \quad t = 0 \\ \text{softmax}(-|R(u,m) - R(u',m')|) & , \quad \text{else} \end{cases}$$

Similarly sample for P(m->u).

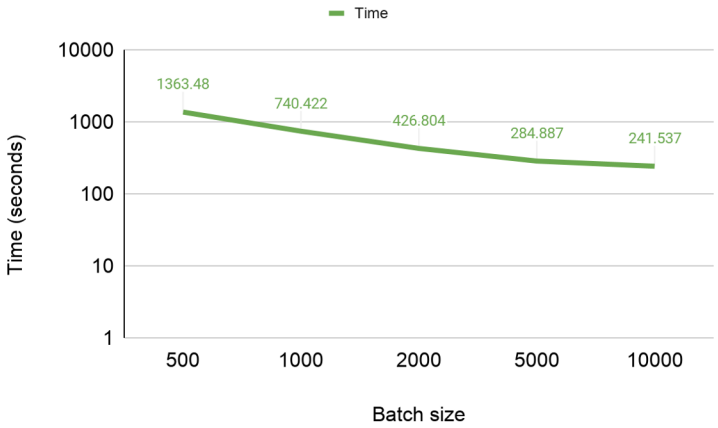
* "metapath2vec: Scalable representation learning for heterogeneous networks." Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. 2017.

Experiment setup

- Methods
  - **Our work**:
    - For **movie representations**: Text only, Graph only, Both text and graph
    - Can change **text embedding** method / **classifier** model…
  - **Other baselines**: SVD, movie2vec
- Metric
  - Mean Absolute Error (**MAE**)
  - Mean Squared Error (**MSE**)
  - **Accuracy**
  - **F1-Score**

Preliminary results

Method: Graph embedding (movies) + MLP (classifier)



Takeaway

- A hybrid recommendation system using text and graph embeddings
- Rating-aware sampling technique
- Evaluation of proposed framework on the dataset