Movie Recommender

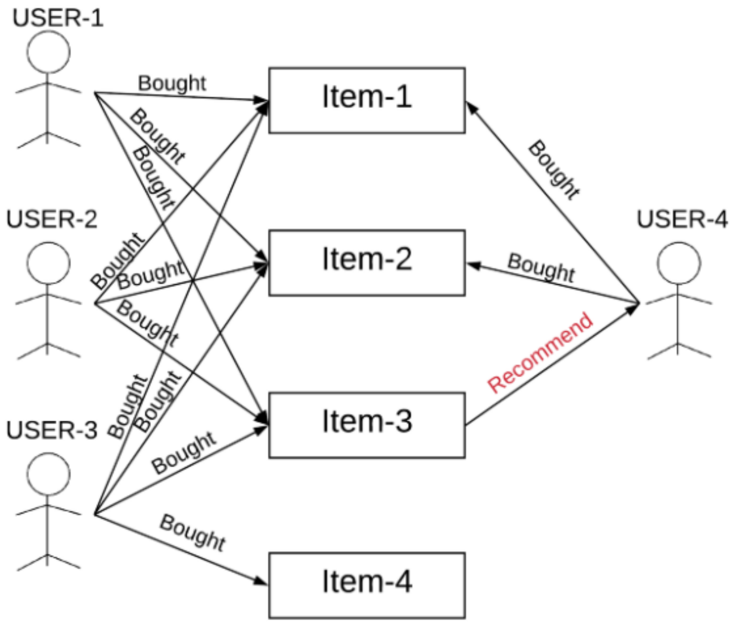
Group member： Zhengqi Dong, Yuntian He

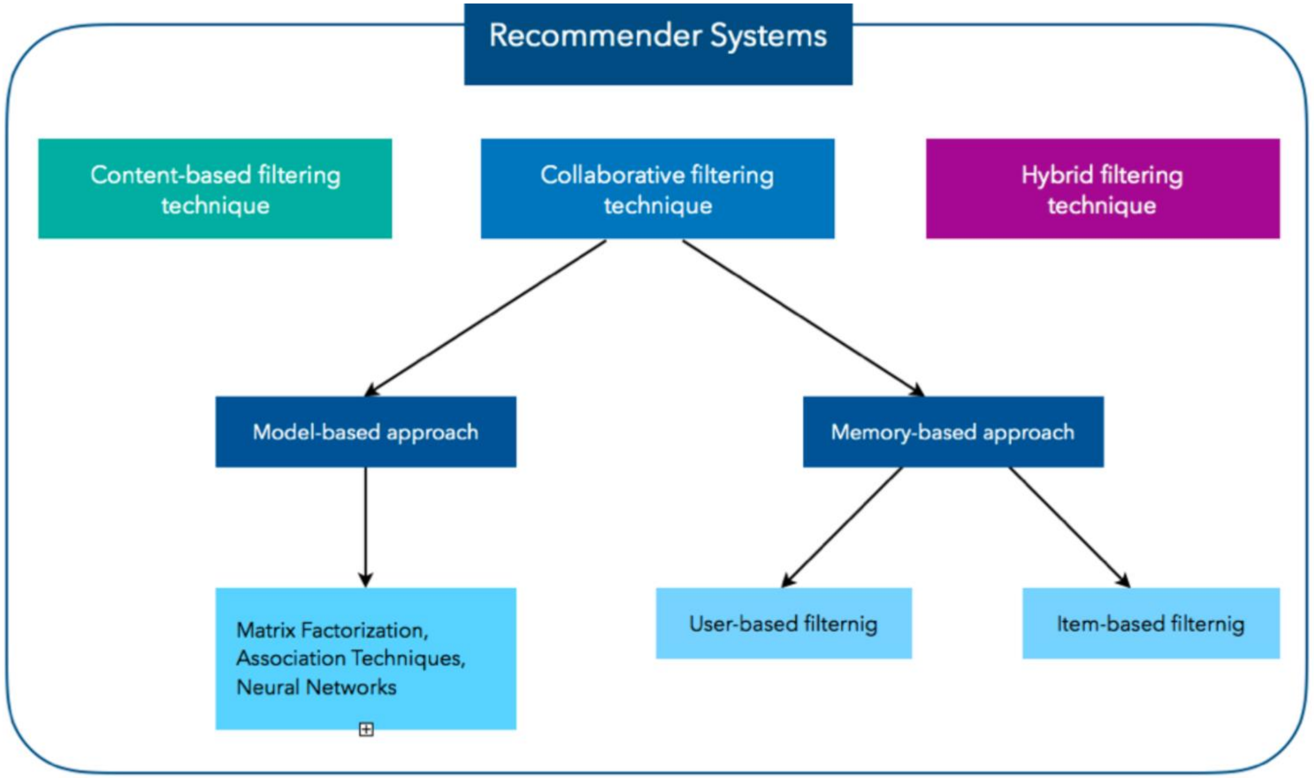
Email: {dong.760, he.1773}@osu.edu

Background

Type of Recommender Systems

- Recommender System:
  - "A recommender system is an information filtering system that seeks to predicts the "rating" or "preference" a user would give to an item."
- Type of Recommender Systems:
  - Content-Based Filtering
  - Collaborative Filtering(CF)
    - Memory-Based Collaborative Filtering, e.g., User-based CF, Item-based CF
    - Model-Based Collaborative Filtering, e.g., Matric factorization, Neural Network
  - Hybrid Filtering





Courtesy: 1) https://d2l.ai/chapter_recommender-systems/recsys-intro.html 2) https://dl.acm.org/doi/pdf/10.1155/2009/421425 3) https://arxiv.org/pdf/1707.07435.pdf, 4) https://humboldt-wi.github.io/blog/research/applied_predictive_modeling_19/causalrecommendersystem/

Dataset description

- **The Movies Dataset** from Kaggle
  - **26M** ratings from **270K** users on **45K** movies
- Content
  - **Text**: Each movie has an overview (a paragraph)
  - **Rating**: A tuple (UserID, MovieID, Rating, Timestamp)
  - **Other Attributes**:
    - Genre: e.g. Action, Animation, Romance, …
    - Credits: (cast, crew)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45466 entries, 0 to 45465
Data columns (total 24 columns):
adult                    45466 non-null object
belongs_to_collection    4494 non-null object
budget                   45466 non-null object
genres                   45466 non-null object
homepage                 7782 non-null object
id                       45466 non-null object
imdb_id                  45449 non-null object
original_language        45455 non-null object
original_title           45466 non-null object
overview                 44512 non-null object
popularity               45461 non-null object
poster_path              45080 non-null object
production_companies     45463 non-null object
production_countries     45463 non-null object
release_date             45379 non-null object
revenue                  45460 non-null float64
runtime                  45203 non-null float64
spoken_languages         45460 non-null object
status                   45379 non-null object
tagline                  20412 non-null object
title                    45460 non-null object
video                    45460 non-null object
vote_average             45460 non-null float64
vote_count               45460 non-null float64
dtypes: float64(4), object(20)
memory usage: 8.3+ MB
```

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

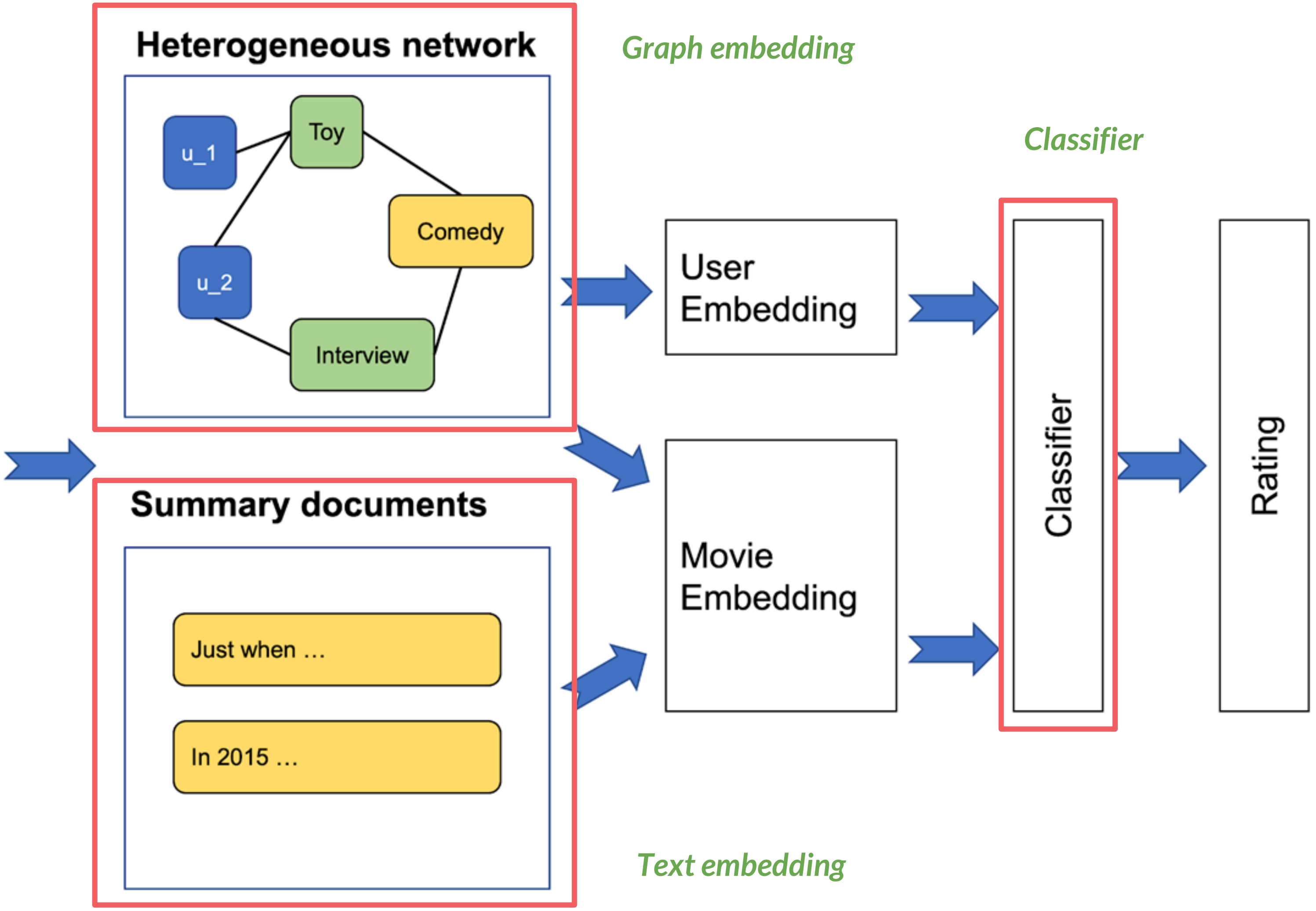
System framework

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, …
…	…	…

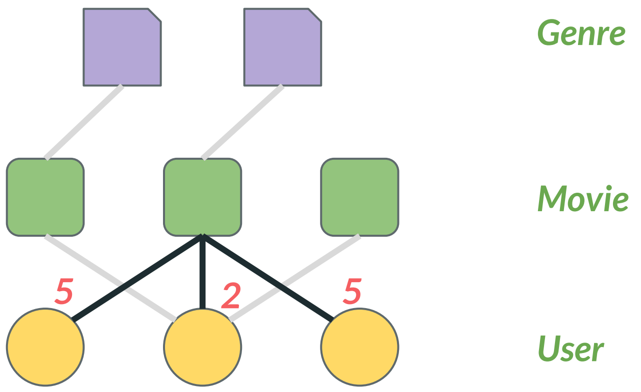


Preprocessing

- 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’

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

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

Similarly sample for P(m->u).

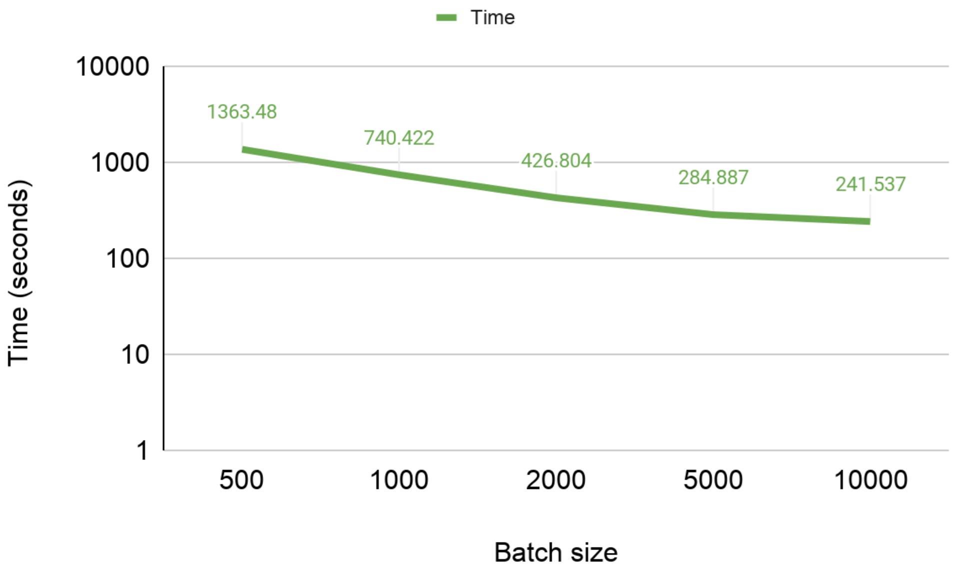
* "metapath2vec: Scalable representation learning for heterogeneous networks." Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. 2017.

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

Thanks! Any Question?

Reference

[1]Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deepbidirectional transformers for language understanding.arXiv preprint arXiv:1810.04805, 2018.

[2]Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. metapath2vec: Scalable representationlearning for heterogeneous networks. InProceedings of the 23rd ACM SIGKDD internationalconference on knowledge discovery and data mining, pages 135–144, 2017.

[3]Yoon Kim.Convolutional neural networks for sentence classification.arXiv preprintarXiv:1408.5882, 2014.

[4]Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. InInternational conference on machine learning, pages 1188–1196, 2014.

[5]Francesco Ricci, Lior Rokach, and Bracha Shapira. Introduction to recommender systemshandbook. InRecommender systems handbook, pages 1–35. Springer, 2011.

[6]Xiaoyuan Su and Taghi M Khoshgoftaar. A survey of collaborative filtering techniques.Advancesin artificial intelligence, 2009, 2009.

[7]Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: Asurvey and new perspectives.ACM Computing Surveys (CSUR), 52(1):1–38, 2019.