Assignment 1: Part A (Question Formation and Exploratory Analysis)

Book Recommendation System

Task Overview

The project poses research questions on personalization of book recommendation systems and gathers, examines, compiles, cleanses, and consolidates massive user interaction datasets to provide answers to such questions. We evaluated data sufficiency through extensive exploratory analytics and broadened our questions in light of the characteristics of the datasets and the limitations we found during the exploration process.

Task Description

Initial Questions of Industrial and Societal Relevance

Primary Research Question: How can machine learning techniques be leveraged to build accurate and stable book recommendation systems that predict user reading preferences based on historical interaction patterns?

Sub-questions:

1. Which recommendation algorithms work best on sparse user-item interaction data in the book domain?

2. How precise and efficient are the different neural architectures, including graph-based, sequential modeling, and collaborative filtering?

3. To improve recommendation performance on implicit feedback data, which model setups and hyper parameters are most effective?

4. What are the best ways to include user behavior sequences and time dynamics in recommendation models?

Industrial Relevance:

- E-commerce sites like Amazon and Barnes & Noble need effective recommendation engines to increase sales and user engagement.
- Digital libraries and reading platforms such as Goodreads and Kindle need tailored content discovery systems.
- Educational institutions can gain from academic resource recommendation systems.
- Publishing houses can enhance their marketing strategies by gaining clearer insights into reading habits.

Societal Relevance:

- Accessibility of information: Facilitates the discovery of appropriate books during the age of information overload
- Reading promotion: Promotes reading habits with tailored recommendations
- Cultural diversity: Encourages varied books and mitigates filter bubble effects
- Educational equity: Helps to pair learners with suitable learning materials

Big Data Definition and Characteristics

Data Sources Identified:

- **Primary Dataset**: Goodbooks - 10k dataset
- **Scale**: ~6 million user-book interactions across 10,000 books and 50,000 users

Big Data Features Satisfied:

1. **Volume:** 6 million interaction records represent a large data volume that requires distributed processing methods.
2. **Variety:** Multi-modal data includes: -
    a. Numerical ratings (1-5 scale)
    b. Temporal timestamps
    c. Textual book metadata (titles, authors, genres)
    d. User demographic information
    e. Review text content
3. **Complexity:** The sparse user-item interaction matrix, with over 99% sparsity, creates computation issues.
4. **Multiple Sources:** Data comes from various user interactions across different time contexts and durations.
5. **Multiple Formats:** The data includes structured numerical data, unstructured text reviews, and categorical metadata.

Data Processing and Assessment

Collation of Data:

1. **Data Gathering:** Obtained Goodbooks-10k dataset from public repositories
2. **Standardization of Format:** Transformed explicit ratings into implicit interaction signals
3. **Data Splitting:** Into training (user-item interaction) and test sets (user prediction targets)
4. **Preprocessing:**
    a. User ID and Item ID normalization
    b. Interaction timestamp processing
    c. Missing value imputation strategies
    d. Standardization of data

Data Quality Assessment:

- **Comprehensive**: ~6M complete interaction records with low missing values
- **Consistency**: Standardization of user and item IDs across datasets
- **Accuracy**: Cross-verified against primary Goodreads data sources
- **Timeliness**: Data reflects modern reading habits (2017–2019)

How Data Addresses Research Questions:

- **Interaction Volume**: Adequate volume for training advanced neural recommendation models
- **Diversity of Users**: 50,000 users offer general behavior pattern representation
- **Item Coverage**: 10,000 books across several genres and levels of popularity
- **Temporal Information**: Timestamps support sequential and temporal modeling methods
- **Implicit Feedback**: Realistic depiction of real-world recommendation situations

Identified Deficiencies and Pitfalls

Data Limitations Discovered:

- **Extreme Sparsity** (>99%):
  - **Impact**: New user/new item cold start issues
  - **Mitigation**: Applied negative sampling methods and integration of auxiliary information
- **Rating Distribution Bias**:
  - **Problem**: Over-representation of high ratings (positive skew)
  - **Solution**: Applied rating normalization and bias correction methods

Question Refinement Based on Data Analysis

Refined Primary Question: How can neural collaborative filtering and sequential recommendation models be optimized to achieve maximum accuracy on sparse implicit feedback data in book recommendation scenarios?

Refined Sub-questions:

1**. Model Architecture Optimization:** What represents the optimal deep neural network designs (depth, attention mechanisms, embedding dimensions) for sparse data book recommendation tasks?

2. **Sequential vs. Collaborative Approaches:** How do self-attention based sequential models (SASRec) perform with respect to graph-based collaborative filtering model (LightGCN, PEAGNN) in book recommendation?

3. **Enhancement Techniques:** How do transformer models and multi-head attention affect the functionality of conventional recommendation systems?

4. **Economy of Computation:** For practical use, how best to strike a compromise between model complexity and training time?

Justification for Refinement:
Focus was directed towards neural architectures due to the following factors:

- sparse implicit feedback data was emphasized
- data availability and scalability were in conflict with model complexity.
- Attention to specified evaluation frameworks and competitive boundaries
- Additional factors such as computing efficiency
- The emphasis on applicability

Next Steps: Analysis Plan

Immediate Analysis Pipeline:

1. **Baseline Model Implementation**:
    - Filtering via Neural Collaboration (NCF)
    - Various forms of matrix factorization
    - Common items and arbitrary baselines
2. **Advanced Model Development**:
    - Utilizing graph-based methods (LightGCN, PEAGNN)
    - Sequential models (SASRec)
    - Temporal models (LT-OCF)
3. **Model Enhancement**:
    - The integration of multi-head attention
    - Adaptation to transformer architecture
    - Ensemble method exploration
4. **Evaluation Framework**:
    - Strategies for cross-validation
    - evaluation of several metrics (HR@K, nDCG@K, F1-score)
    - analysis of computational efficiency

Tools and Technologies Planned:

- **Deep Learning Frameworks**: PyTorch, TensorFlow
- **Data Processing**: Pandas, NumPy, Scipy
- **Visualization**: Matplotlib, Seaborn
- **Evaluation**: Scikit-learn, custom evaluation metrics
- **Computational Infrastructure**: GPU-accelerated training environments

Backup Question and Alternative Data Sources

Backup Question 1: Cross-Domain Recommendation Transfer Research Question: How effectively can recommendation models trained on book data transfer to other media domains (movies, music, articles)?

Other Resources:

- **MovieLens Dataset**: 25M movie ratings for cross-domain evaluation
- **Amazon Product Reviews**: Multi-category product interactions
- **Last.fm Music Dataset**: Music listening behavior data
- **News Article Interaction Data:** Reading pattern analysis

Backup Question 2: Recommendation Fairness and Bias Analysis Research Question: How do different recommendation algorithms perform across demographic groups and item popularity distributions in book recommendations?

Other Resources:

- **LibraryThing Dataset**: Broader demographic information
- **Open Library Data**: Comprehensive book metadata
- **Academic Paper Citation Networks**: Scholarly recommendation scenarios
- **Social Reading Platforms**: Community-based reading data

Backup Question 3: Content-Enhanced Recommendation Systems Research Question: How can textual content (book summaries, reviews, author information) improve collaborative filtering performance?

Other Resources:

- **Google Books API**: Rich book metadata and content
- **Wikipedia Book Pages**: Comprehensive book information
- **Publisher Databases**: Genre and category hierarchies
- **Academic Literature Databases**: Citation and topic modeling data

Reference

He, X., Liao, L., Zhang, H., Nie, L., Hu, X. and Chua, T.-S. (2017). Neural Collaborative Filtering. *arXiv:1708.05031*. Available at: https://arxiv.org/abs/1708.05031.

Kang, W.-C. and McAuley, J. (2018). Self-Attentive Sequential Recommendation. *arXiv:1808.09781*. Available at: https://arxiv.org/abs/1808.09781.

Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W. and Jiang, P. (2019). BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer. *arxiv.org*. doi:https://doi.org/10.48550/arXiv.1904.06690.

Wang, X., He, X., Wang, M., Feng, F. and Chua, T.-S. (2019). Neural Graph Collaborative Filtering. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. doi:https://doi.org/10.1145/3331184.3331267.

He, X., Deng, K., Wang, X., Li, Y., Zhang, Y. and Wang, M. (2020). LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. *arXiv:2002.02126*. Available at: https://arxiv.org/abs/2002.02126.

Umer, A.M., Han, Z., Arumugaswamy, S., Khan, R.A., Weber, T., Qiu, T., Shen, H., Liu, Y. and Kleinsteuber, M. (2020). *Metapath- and Entity-aware Graph Neural Network for Recommendation*. arXiv.org. Available at: https://arxiv.org/abs/2010.11793.

Wang, Y., Ma, W., Zhang, M., Liu, Y. and Ma, S. (2022). A Survey on the Fairness of Recommender Systems. *ACM Transactions on Information Systems*, 41(3). doi:https://doi.org/10.1145/3547333.

Anastasiia Klimashevskaia, Dietmar Jannach, Elahi, M. and Trattner, C. (2024). A survey on popularity bias in recommender systems. *User Modeling and User-Adapted Interaction*. doi:https://doi.org/10.1007/s11257-024-09406-0.